

Asymptotiques de fonctionnelles d'arbres aléatoires et de graphes denses aléatoires

Marion Sciauveau

▶ **To cite this version:**

Marion Sciauveau. Asymptotiques de fonctionnelles d'arbres aléatoires et de graphes denses aléatoires. Algèbres d'opérateurs [math.OA]. Université Paris-Est, 2018. Français. NNT : 2018PESC1127 . tel-02084023

HAL Id: tel-02084023
https://pastel.hal.science/tel-02084023

Submitted on 29 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS-EST

L'École Doctorale Mathématiques et Sciences et Technologies de l'information et de la Communication (MSTIC)

Thèse de doctorat

Discipline : Mathématiques

présentée par

Marion SCIAUVEAU

Asymptotiques de fonctionnelles d'arbres aléatoires et de graphes denses aléatoires

---

Thèse co-dirigée par Jean-François Delmas et Jean-Stéphane Dhersin

préparée au CERMICS, École des Ponts ParisTech

Thèse soutenue le 14 novembre 2018 devant le Jury composé de :

M. Pierre Calka	Examinateur	Université de Rouen
M. Philippe Chassaing	Président	Université Henri Poincaré
Mme Brigitte Chauvin	Examinatrice	Université Paris-Saclay
M. Jean-François Delmas	Directeur de Thèse	École des Ponts ParisTech
M. Jean-Stéphane Dhersin	Directeur de Thèse	Université Paris 13
M. Jean-François Marckert	Rapporteur	Université de Bordeaux

Rapporteurs:

Mme Christina Goldschmidt	Université d'Oxford
M. Jean-François Marckert	Université de Bordeaux

REMERCIEMENTS

C'est au moment d'écrire ces remerciements que je réalise le chemin parcouru en trois ans ! Une thèse, c'est une aventure ponctuée de doutes et d'échecs mais aussi et surtout de bonheurs et réussites. C'est également de belles rencontres.

Je tenais tout d'abord à remercier mes deux directeurs de thèse, Jean-François Delmas et Jean-Stéphane Dhersin. Merci de m'avoir fait confiance en décidant d'encadrer ma thèse. Grâce à vous, j'ai pu découvrir la recherche dans les meilleures conditions qui soient. Vos conseils et vos connaissances m'ont été précieux. Merci pour le temps que vous m'avez consacré. Je n'oublierai pas nos réunions, où je ressortais épuisée mais en ayant appris tellement de choses nouvelles. Jean-François, merci d'avoir toujours été disponible pour répondre à mes questions mathématiques et parfois existentielles. Vous êtes un directeur exceptionnel et ces quelques mots ne suffiront pas à dire à quel point je vous suis reconnaissante de votre engagement, de votre soutien et de votre patience.

Mes remerciements vont ensuite à mes deux rapporteurs Christina Goldschmidt et Jean-François Marckert pour leur relecture attentive et leurs commentaires constructifs qui m'ont permis d'améliorer mon manuscrit. Je remercie également Pierre Calka, Philippe Chassaing et Brigitte Chauvin d'avoir gentiment accepté de faire partie du jury. J'en suis honorée. Pierre, ta présence dans mon jury me touche particulièrement. Nous sommes arrivés la même année à l'université de Rouen, toi en tant que professeur, moi en tant qu'étudiante. Tu m'as donné le goût des probabilités, j'en ai fait le domaine de ma thèse. C'est grâce à toi que j'ai pu travailler avec Jean-Stéphane et Jean-François, merci pour ta confiance.

Mes trois années de thèse ont été particulièrement agréables grâce au cadre de travail idéal du Cermics. Merci à tous les chercheurs du laboratoire, pour leur gentillesse et leur disponibilité. Je pense notamment à Jean-Philippe Chancelier, Bernard Lapeyre, Frédéric Meunier et Pierre-André Zitt avec qui j'ai pu découvrir le monde des graphons durant nos groupes de lecture. Merci Isabelle pour toutes tes attentions et ton efficacité redoutable. Je tenais également à remercier Fatna pour sa gentillesse. Encore merci à toutes les deux, vous m'avez été d'une grande aide durant ces trois années.

Je voudrais maintenant remercier tous les doctorants, stagiaires et post-doctorants que j'ai pu côtoyer durant ma thèse. Je pense en particulier à la team du deuxième étage : Adèle, Alexandre (mon petit matara), Étienne, Henri (mon voisin de bureau toujours au top), Gustave, Laurent, merci pour votre bonne humeur, pour toutes nos discussions et fous rires qui ont rendu cette aventure plus folle ! Une pensée également pour les petits derniers, Adel, Benoît, Danil, Mouad, Oumaima, Victor et William. Bon courage pour la suite ! Je n'oublie pas non plus, Adrien, Amina, Florent, François, Fred, Grégoire, Julien, Karol, Laura, Lingling,

Marc, Rafaël, Sami et Pierre-Loïk.

Merci à Amélie et Lucie pour nos discussions, leur soutien avant la soutenance et la répétition à Tours.

J'ai eu la chance de pouvoir enseigner en tant que chargée de TDs à l'université Paris 13 au sein de l'institut Galilée. Je remercie donc l'ensemble des membres du département de mathématiques pour leur accueil et plus particulièrement l'équipe de probabilités dont j'ai pu suivre le séminaire chaque mercredi. Merci aux responsables de cours, Isabelle Gaudron, Bénédicte Haas et Laurent Tournier, cela a été un plaisir de découvrir l'enseignement à vos côtés. Je voudrais également remercier la secrétaire du laboratoire, Isabelle Barbotin pour son aide. Enfin merci aux doctorants de Paris 13, qui m'ont souvent croisée en coup de vent entre deux TDs. Je pense en particulier à Anna-Laura, Delphin, Irène et Pierre.

Je remercie la région Île-de-France d'avoir financé ma thèse, me permettant ainsi de travailler dans d'excellentes conditions. Je tenais à remercier Dominique Wetzel de l'équipe du DIM pour s'être si bien occupée de nous.

Je n'aurais jamais eu la chance ou même l'idée de faire une thèse sans l'implication et la ténacité de certains chercheurs de l'université de Rouen. Je me dois donc de remercier Olivier Benois et encore une fois Pierre Calka de m'avoir toujours poussée à me dépasser. Non seulement vous avez toujours été présents pour répondre à mes nombreuses questions mathématiques mais vous avez également toujours su trouver les mots justes pour me redonner confiance en moi. Je vous en suis particulièrement reconnaissante. Je remercie également Jean-Baptiste Bardet, Gaëlle Chagny, Antoine Channarond, Olivier Guibé, Paul Lescot, Mustapha Mourragui et Simon Raulot pour leurs enseignements de qualité.

Il est maintenant temps de remercier mes amis sur lesquels j'ai toujours pu compter et dont le soutien m'a été essentiel. Merci à Aurélie, Benji et Flo, mes compères de l'université : nos parties de cartes endiablées, nos courses de stylo, nos blagues en tout genre et nos fous rires resteront inoubliables . Merci à Clémentine, Déborah, Léopoldine, Sophie et Thomas pour avoir rendu l'année de préparation à l'agrégation mémorable. Enfin un énorme merci à Zoë ; que de chemin parcouru depuis le CP ! Merci d'avoir toujours su me remotiver alors même que tu ne comprenais pas vraiment l'idée que j'avais eu de faire une thèse. Nos rendez-vous Starbucks le week-end étaient indispensables.

Je terminerai en remerciant mes proches pour leur soutien indéfectible. Merci Florence, tu es une marraine au top. Merci à mes cousins Brice et Mathilde que j'adore par dessus tout. Merci à ma mamie Colette de s'être déplacée pour assister à ma soutenance. J'aurais aimé qu'en ce jour si particulier, mes papis, Maurice et Paul et ma mamie, Marguerite aient pu t'accompagner. Merci également à marrain, Jean-Marc, Anne et Serge. Un tendre merci à Thibault. Tu as su me soutenir et me comprendre tout au long de cette aventure. Les moments passés ensemble m'ont été précieux. Enfin un énorme merci à mes parents et mon frère Nicolas pour leur présence et leur amour. Merci de m'avoir permis de faire le métier dont je rêvais en me soutenant toujours même quand mes choix vous paraissaient étranges. Merci pour tout ce que je vous dois.

Résumé

L'objectif de cette thèse est l'étude des approximations et des vitesses de convergence pour des fonctionnelles de grands graphes discrets vers leurs limites continues. Nous envisageons deux cas de graphes discrets : des arbres (i.e. des graphes connexes et sans cycles) et des graphes finis, simples et denses.

Dans le premier cas, on considère des fonctionnelles additives sur deux modèles d'arbres aléatoires : le modèle de Catalan sur les arbres binaires (où un arbre est choisi avec probabilité uniforme sur l'ensemble des arbres binaires complets ayant un nombre de nœuds donné) et les arbres simplement générés (et plus particulièrement les arbres de Galton-Watson conditionnés par leur nombre de nœuds).

Les résultats asymptotiques reposent sur les limites d'échelle d'arbres de Galton-Watson conditionnés. En effet, lorsque la loi de reproduction est critique et de variance finie (ce qui est le cas des arbres binaires de Catalan), les arbres de Galton-Watson conditionnés à avoir un grand nombre de nœuds convergent vers l'arbre brownien continu qui est un arbre réel continu qui peut être codé par l'excursion brownienne normalisée. Par ailleurs, les arbres binaires sous le modèle de Catalan peuvent être construits comme des sous-arbres de l'arbre brownien continu. Ce plongement permet d'obtenir des convergences presque-sûres de fonctionnelles. Plus généralement, lorsque la loi de reproduction est critique et appartient au domaine d'attraction d'une loi stable, les arbres de Galton-Watson conditionnés à avoir un grand nombre de nœuds convergent vers des arbres de Lévy stables, ce qui permet d'obtenir le comportement asymptotique des fonctionnelles additives pour certains arbres simplement générés.

Dans le second cas, on s'intéresse à la convergence de la fonction de répartition empirique des degrés ainsi qu'aux densités d'homomorphismes de suites de graphes finis, simples et denses. Une suite de graphes finis, simples, denses converge si la suite réelle des densités d'homomorphismes associées converge pour tout graphe fini simple. La limite d'une telle suite de graphes peut être décrite par une fonction symétrique mesurable appelée graphon.

Étant donné un graphon, on peut construire par échantillonnage, une suite de graphes qui converge vers ce graphon. Nous avons étudié le comportement asymptotique de la fonction de répartition empirique des degrés et de mesures aléatoires construites à partir des densités d'homomorphismes associées à cette suite particulière de graphes denses.

Mots-clés : Arbre aléatoire binaire, arbre de Galton-Watson conditionné, fonctionnelle de coût, excursion brownienne, arbre réel continu, graphon, graphe dense, densité d'homomorphisme, fonction de distribution des degrés, mesure aléatoire.

Classification AMS : 05C05, 05C07, 05C80, 60C05, 60F05, 60F17, 60G57, 60J80.

Abstract

The aim of this thesis is the study of approximations and rates of convergence for functionals of large discrete graphs towards their limits. We contemplate two cases of discrete graphs: trees (i.e. connected graphs without cycles) and dense simple finite graphs.

In the first case, we consider additive functionals for two models of random trees: the Catalan model for binary trees (where a tree is chosen uniformly at random from the set of full binary trees with a given number of nodes) and the simply generated trees (and more particularly the Galton-Watson trees conditioned by their number of nodes).

Asymptotic results are based on scaling limits of conditioned Galton-Watson trees. Indeed, when the offspring distribution is critical and with finite variance (that is the case of Catalan binary trees), the Galton-Watson trees conditioned to have a large number of nodes converge towards the Brownian continuum tree which is a real tree coded which can be coded by the normalized Brownian excursion. Furthermore, binary trees under the Catalan model can be built as sub-trees of the Brownian continuum tree. This embedding makes it possible to obtain almost sure convergences of functionals. More generally, when the offspring distribution is critical and belongs to the domain of attraction of a stable distribution, the Galton-Watson trees conditioned to have a large number of nodes converge to stable Lévy trees giving the asymptotic behaviour of additive functionals for some simply generated trees.

In the second case, we are interested in the convergence of the empirical cumulative distribution of degrees and the homomorphism densities of sequences of dense simple finite graphs. A sequence of dense simple finite graphs converges if the real sequence of associated homomorphism densities converges for all simple finite graph. The limit of such a sequence of dense graphs can be described as a symmetric measurable function called graphon. Given a graphon, we can construct by sampling, a sequence of graphs which converges towards this graphon. We have studied the asymptotic behaviour of the empirical cumulative distribution of degrees and random measures built from homomorphism densities associated to this special sequence of dense graphs.

Keywords: Random binary tree, conditioned Galton-Watson tree, cost functional, Brownian excursion, continuum random tree, graphon, dense graph, homomorphism density, cumulative distribution function of degrees, random measure.

AMS classification: 05C05, 05C07, 05C80, 60C05, 60F05, 60F17, 60G57, 60J80.

TABLE DES MATIÈRES

Introduction et présentation des résultats **11**

I Présentation des résultats **17**

1 Fonctionnelles additives sur les arbres aléatoires **19**
1.1 Arbres discrets . . . 19
1.1.1 Notions sur les arbres discrets . . . 19
1.1.2 Arbres binaires . . . 21
1.2 Arbres de Galton-Watson . . . 23
1.2.1 Définitions . . . 23
1.2.2 Arbres de Galton-Watson conditionnés et arbres simplement générés . 24
1.2.3 Convergence du processus de contour des arbres de Galton-Watson conditionnés . . . 25
1.3 Arbres réels . . . 27
1.3.1 Définitions . . . 27
1.3.2 Arbre brownien . . . 29
1.3.3 Construction d'arbres binaires à partir de l'arbre continu brownien . . 29
1.4 Fonctionnelles de coût et fonctions péage . . . 31
1.4.1 Définition . . . 32
1.4.2 Transition de phase . . . 32
1.4.3 Fonctionnelles de coût globales . . . 33
1.5 Résultats sur les fonctionnelles globales du Chapitre 3 . . . 35
1.5.1 Résultats pour les arbres binaires sous le modèle de Catalan . . . 36
1.5.2 Résultats pour les arbres simplement générés . . . 38
1.5.3 Extensions possibles . . . 40

2 Fonctionnelles de graphes aléatoires échantillonnés à partir d'un graphon **41**
2.1 Graphes . . . 41
2.1.1 Notions sur les graphes . . . 41
2.1.2 Graphes aléatoires . . . 42
2.1.3 Heuristique de la convergence d'une suite de graphes aléatoires . . . 42
2.2 Des graphes aux graphons . . . 43
2.2.1 Densités d'homomorphismes . . . 43
2.2.2 Graphons . . . 44
2.2.3 Convergence d'une suite de graphes denses . . . 45
2.3 Échantillonnage de graphes aléatoires à partir d'un graphon . . . 47

2.4 Résultats du Chapitre 4 . . . 50
2.4.1 Fonction de répartition empirique des degrés . . . 50
2.4.2 Densités d'homomorphismes de graphes partiellement étiquetés . . . 53
2.4.3 Extensions possibles . . . 55

II Résultats **57**

3 Fonctionnelles de coût de grands arbres aléatoires (uniformes et simplement générés) **59**
3.1 Introduction . . . 59
3.1.1 A finite measure indexed by a tree . . . 60
3.1.2 Additive functionals and toll functions for binary trees . . . 61
3.1.3 Main results on the asymptotics of additive functionals in the Catalan model . . . 62
3.1.4 Main results on the asymptotics of additive functionals for simply generated trees . . . 64
3.1.5 Organization of the paper . . . 65
3.2 Notations . . . 65
3.2.1 Ordered rooted discrete trees . . . 65
3.2.2 Real trees . . . 66
3.2.3 The Brownian continuum random tree $\mathcal{T}$ . . . 67
3.2.4 The discrete binary tree from the Brownian tree . . . 67
3.2.5 Simply generated random tree . . . 68
3.3 Catalan model . . . 69
3.4 Simply generated trees model . . . 71
3.4.1 Contour process . . . 71
3.4.2 Convergence of contour processes . . . 72
3.4.3 Main result . . . 73
3.4.4 Proof of Corollary 3.15 . . . 74
3.5 Preliminary Lemmas . . . 76
3.6 Proof of Theorem 3.4 . . . 81
3.7 Proof of Proposition 3.10 . . . 82
3.7.1 A preliminary convergence in distribution . . . 82
3.7.2 Proof of Proposition 3.10 . . . 84
3.8 Appendix . . . 86
3.8.1 Upper bounds for moments of the cost functional . . . 86
3.8.2 A lemma for binomial random variables . . . 86
3.8.3 Some results on the Gamma function . . . 87
3.8.4 Elementary computations on the branch length of $\mathcal{T}_{[n]}$ . . . 88
3.8.5 A deterministic representation formula . . . 89
3.8.6 Proof of the first part of Lemma 3.17 (finiteness of Z_β^H and (3.23)) . . . 91

4 Asymptotiques pour la fonction de répartition empirique des degrés et les densités d'homomorphismes de graphes aléatoires échantillonnés à partir d'un graphon **95**
4.1 Introduction . . . 95
4.1.1 Convergence of CDF of empirical degrees for large random graphs . . . 96
4.1.2 Convergence of sequence of dense graphs towards graphons . . . 97
4.1.3 Asymptotics for homomorphism densities of partially labeled graphs for large random graphs . . . 99
4.1.4 Organization of the paper . . . 101

4.2 Definitions . . . 101
4.2.1 First notation . . . 101
4.2.2 Graph homomorphisms . . . 102
4.2.3 Graphons . . . 105
4.2.4 W-random graphs . . . 106
4.3 Asymptotics for homomorphism densities of sampling partially labeled graphs from a graphon . . . 107
4.3.1 Random measures associated to a graphon . . . 107
4.3.2 Invariance principle and its fluctuations . . . 109
4.4 A preliminary result . . . 112
4.5 Proof of Theorem 4.9 . . . 115
4.6 Proof of Theorem 4.11 . . . 116
4.7 Asymptotics for the empirical degrees cumulative distribution function of the degrees . . . 118
4.8 Preliminary results for the empirical cdf of the degrees . . . 119
4.8.1 Estimates for the first moment of the empirical cdf . . . 119
4.8.2 Estimates for the second moment of the empirical cdf . . . 122
4.9 Proof of Theorem 4.23 . . . 125
4.10 Appendix A: Preliminary results for the CDF of binomial distributions . . . 130
4.11 Appendix B: Proof of Proposition 4.28 . . . 137
4.11.1 A preliminary result . . . 137
4.11.2 Proof of Proposition 4.28 . . . 140

Références bibliographiques **141**

INTRODUCTION ET PRÉSENTATION DES RÉSULTATS

Dans cette thèse, nous nous intéressons à l'étude asymptotique de fonctionnelles additives sur de grands graphes aléatoires. Les résultats obtenus reposent de manière déterminante sur le passage d'objets discrets, à savoir les arbres et les graphes aléatoires, vers leurs objets limites respectifs, les arbres réels continus et les graphons.

En guise d'introduction générale à cette thèse, nous avons décidé de faire un bref rappel historique sur la naissance et le développement des arbres et graphes aléatoires, afin de motiver nos travaux. Celui-ci ne prétend bien entendu pas être exhaustif. Nous donnons ensuite un résumé de nos travaux.

Historique et motivations sur les arbres aléatoires

Les arbres, au sens mathématiques, sont des graphes connexes et sans cycles. D'après Knuth [126], le concept d'arbres a été défini pour la première fois de manière mathématique en 1847 par Kirchhoff en lien avec les réseaux électriques. Le mot « arbre » à proprement parler apparaît quant à lui, une décennie plus tard, dans une série de papiers de Cayley [43] qui s'intéresse à l'énumération des arbres. Les arbres sont des structures fondamentales qui apparaissent de manière naturelle dans de nombreux domaines.

Arbres en informatique

En informatique, les arbres sont vus comme des structures de données récursives : un « nœud » appelé « racine » et ses « enfants », qui sont eux-mêmes des arbres. La suite d'ouvrages « The art of computer programming » de Knuth [126, 127, 128], publiée au tournant des années 60 et 70, constitue les fondations de l'analyse algorithmique et présente les arbres comme des structures de données. En effet, ils permettent de classer, stocker et chercher des données de manière efficace.

Les arbres binaires de recherche sont un exemple d'arbres binaires (arbres où chaque nœud possède au plus deux enfants) auxquels on associe des données, aussi appelées clés, à ses nœuds internes. Ils ont été introduits dans le début des années 1960. Ces arbres sont associés à des algorithmes de tri, tels que l'algorithme de tri rapide (« quicksort algorithm » en anglais). D'un point de vue combinatoriste, les arbres binaires de recherche sont simplement des arbres binaires.

A partir des années 90, l'étude des algorithmes en lien avec les arbres de recherche est développée dans un cadre probabiliste. Le lecteur pourra se référer aux ouvrages de Mahmoud [142] en 1992, Devroye [58] en 1998, Drmota [68] en 2009, Sedgewick et Flajolet [177] en 2013 et Chauvin, Clément et Gardy [47] en 2018 qui y développent des méthodes à la fois combinatoires et probabilistes.

Arbres en biologie

Les arbres jouent également un rôle très important en biologie. Ils interviennent ainsi en phylogénie où ils permettent de classer des espèces et de mettre en évidence leurs relations, voir Semple et Steel [178]. Les arbres phylogénétiques ne sont rien d'autres que des arbres binaires enracinés dont les feuilles, correspondant aux espèces, sont étiquetées. Quand le nombre d'espèces augmentent, il est intéressant de regarder la forme des arbres afin de mieux comprendre leur structure. Pour ce faire, différents indices, ne dépendant que de la forme des arbres, ont été considérées (voir voir Shao et Sokal [179], Heard [104], Kirkpatrick et Slatkin [123], Mooers et Heard [149], Felsenstein [82], chapitre 33).

Les arbres en biologie sont aussi présents au travers des arbres de Galton-Watson. Les processus de Galton-Watson ont été introduits au milieu du 19ième siècle afin d'étudier le problème d'extinction des noms de famille noble. En 1845, Bienaymé est le premier à s'intéresser à ce problème et énonce une solution formulée dans un cadre non mathématique. En 1875, Galton et Watson ont donné une preuve mathématique de ce problème. Celle-ci n'était cependant pas exacte puisqu'ils concluaient à l'extinction presque sûre de tous les noms de famille. La première preuve complète et exacte a été donnée en 1930 par Steffensen. Le lecteur pourra se référer aux ouvrages de Kendall [120] et Bacaër [16] pour plus de références bibliographiques.

Arbres aléatoires

C'est à partir des années $1980-1990$ que l'étude probabiliste des arbres aléatoires s'étend considérablement grâce au développement de nouvelles méthodes probabilistes telles que les martingales, les différentes notions de convergence, la méthode de contraction, les théorèmes limites ou encore les processus de branchement et marches aléatoires. Le formalisme de Neveu sur les arbres est ainsi explicité en 1986 [153]. Parmi les arbres étudiés, certains appartiennent à des classes combinatoires telles que les arbres binaires, les arbres planaires ou encore les arbres étiquetés ou non étiquetés, enracinés ou non enracinés. On associe à ces classes d'arbres, le modèle de probabilité naturel consistant à choisir un arbre uniformément au hasard parmi tous les arbres d'une même classe et ayant le même nombre de nœuds, voir Drmota [68]. On peut également considérer des arbres qui évoluent selon un processus stochastique. Ce sont les arbres de branchement dont les arbres simplement générés et les arbres de Galton-Watson sont deux exemples. Les arbres simplement générés introduits pas Meir et Moon [144] en 1978 sont des arbres enracinés auxquels on associe des poids. Il est intéressant de noter que plusieurs classes combinatoires d'arbres peuvent être réalisés comme des arbres simplement générés. Ils peuvent également être vus comme une généralisation des arbres de Galton-Watson (au sens où les arbres de Galton-Watson conditionnés par leur taille totale sont des cas particuliers d'arbres simplement générés). L'utilisation des processus de branchement et des marches aléatoires a permis l'étude asymptotique de nombreuses caractéristiques de ces arbres telles que la hauteur, le nombre total de sommets, la longueur du chemin totale ou encore le nombre de sommets de degré fixé, voir Devroye [55, 56, 59], Drmota [68] et Holmgren et Janson [109], section 13.

Arbres réels continus

Les arbres réels ont été étudiés depuis très longtemps à des fins algébriques et géométriques (voir par exemple Dress, Moulton et Terhalle [67]). Les travaux d'Evans, Pitman et Winter [79] et d'Evans [78] ont donné un cadre formel aux arbres réels.

Une avancée majeure dans l'étude des arbres aléatoires a été l'invention de l'arbre continu brownien par Aldous [8] en 1991. En effet, Aldous a eu l'idée de ne plus simplement s'intéresser

à la convergence de certaines statistiques des arbres mais à la convergence des arbres eux-mêmes. Il a ainsi considéré des limites d'échelle de plusieurs classes d'arbres conditionnés à être grands. Il a montré qu'un arbre de Galton-Watson dont la loi de reproduction est une loi de Poisson de paramètre 1 et conditionné à avoir n nœuds, converge quand n tend vers l'infini vers ce qui a été appelé l'arbre continu brownien. Dans [9, 10], Aldous a également montré que l'arbre continu brownien pouvait être codé par l'excursion brownienne normalisée (de la même manière que les arbres discrets sont codés par leur fonction de contour) et que le processus de contour renormalisé d'un arbre de Galton-Watson, dont la loi de reproduction est critique et de variance finie, et conditionné à avoir n nœuds, converge quand n tend vers l'infini, vers l'excursion brownienne normalisée.

Grâce aux caractéristiques des arbres aléatoires continus, beaucoup de propriétés combinatoires des arbres discrets associés à des processus de branchement ont pu être expliquées. Les arbres continus apparaissent aussi naturellement comme limite d'échelle des processus de branchement.

Les arbres de Lévy, introduits pas Le Gall et Le Jan [136] en 1998, voir aussi la monographie de Duquesne et Le Gall [70], sont une généralisation de l'arbre continu brownien d'Aldous et apparaissent de manière naturelle comme limite d'échelle des arbres de Galton-Watson. Duquesne [69] a montré que la fonction de contour, correctement renormalisée d'un arbre de Galton-Watson dont la loi de reproduction est dans le domaine d'attraction d'une loi stable, et conditionné à avoir n sommets, converge en loi, lorsque n tend vers l'infini, vers le processus de hauteur codant l'arbre de Lévy stable.

Historique et motivations sur les graphes aléatoires

La théorie des graphes est un sujet ancien, dont l'invention serait due à Euler en 1736 pour résoudre le célèbre problème des ponts de Königsberg.

Graphes aléatoires

Les premiers modèles de graphes aléatoires ont été introduits en $1959-1960$ par Erdős et Rényi [74, 75, 76, 77] et Gilbert [99] dans leurs papiers fondateurs (voir aussi Bollobás [30] et Durrett [72], Chapitre 2 pour des références sur le sujet). Ce modèle de graphes aléatoires est dit « homogène » au sens où les degrés de chaque sommet ont même distribution : ils suivent une loi binomiale. Ce modèle est à la fois simple à décrire et à étudier ce qui en fait un excellent outil pour démontrer des propriétés intéressantes de graphes. Ainsi, il possède une transition de phase dans la taille de sa composante connexe maximale quand son paramètre varie, voir par exemple Bollobás [29], Durrett [72], chapitre 2 ou van der Hofstad [187], chapitre 4.

Quelques années plus tard, en 1999, initié par les observations de Faloutsos, Faloutsos et Faloutsos [80] un grand intérêt a été porté à l'étude des réseaux qui nous entourent dans la vie réelle. Ces réseaux sont très grands et dans la plupart des cas, il est difficile de les décrire en détail et de donner un modèle à la fois réaliste et exploitable. Parmi ces réseaux figurent Internet, les réseaux sociaux, les réseaux de citations dans les articles de recherche scientifique (étudiés par Price [164] en 1965), les réseaux de télécommunication, les réseaux biologiques (neurones, cellules, interactions entre des protéines, ...) ou encore les réseaux de collaborations. Leurs études ont mis en évidence des propriétés différentes de celles des graphes d'Erdős-Rényi. Ce sont des réseaux dits « petits mondes » (« small worlds » en anglais), i.e. les distances typiques entre deux sommets sont petites. Ces réseaux ont été introduits par Watts et Strogatz [192] en 1998 (voir aussi Durrett [72], chapitre 5). Les réseaux réels sont également des réseaux invariant d'échelle (« scale-free networks » en anglais), c'est à dire qu'ils ont des degrés qui suivent des lois de puissance (« power laws » en anglais). Les sommets peuvent être « nés » à des moments différents engendrant ainsi des sommets anciens et récents ayant

des propriétés très différentes. On dit que ces graphes sont « inhomogènes ». De nombreux modèles de graphes aléatoires, plus complexes que le modèle d'Erdős-Rényi, ont des propriétés qui se rapprochent de celles observées sur les réseaux réels, voir Newman [158, 154], Albert et Barabási [6], Bollobás [29], Janson, Łuczak et Ruciński [117], Durrett [72] et van der Hofstad [187]. Citons par exemple :

- le modèle de configuration (modèle de graphes aléatoires où les degrés des sommets sont fixés) introduit en 1980 par Bollobás [28] dans le contexte des graphes réguliers, voir aussi Molloy et Reed [147], Newman, Strogatz et Watts, [156, 157], Durrett [72], chapitre 3, van der Hofstad [187], chapitre 7 ;
- le modèle d'attachement préférentiel (modèle de graphes aléatoires dynamique, construits de manière récursive) introduit par Barabási et Albert [17], voir aussi Durrett [72], chapitre 4, Bollobás [29] ou van der Hofstad [187], chapitre 8 ;
- le modèle de graphes inhomogènes défini par Bollobás, Janson et Riordan [31], voir aussi van der Hofstad [187], chapitre 6 ;
- le modèle de graphes exponentiels (modèle statistique qui donne une plus grande probabilité aux graphes qui correspondent le mieux aux caractéristiques observées) introduit par Holland et Leinhardt [107] en 1981, voir Strauss [180], Frank et Strauss [94], Anderson, Wasserman et Crouch [14], Park et Newman [162] ou encore Robins, Pattison, Kalish et Lusher [170].

Le livre de Newman, Barabási et Watts [155] de 2006 est une compilation d'articles qui ont marqué le développement des graphes aléatoires. A noter, que beaucoup d'articles sur les graphes aléatoires ont été produits dans d'autres domaines que les mathématiques. Ainsi, la littérature à ce sujet est très riche en informatique, en physiques ainsi qu'en sciences naturelles.

Limites de graphes aléatoires

L'étude des limites de graphes a commencé au début des années 2000, notamment menée par le pôle recherche de Microsoft. De la même manière que pour les arbres aléatoires, l'objectif était de ne plus seulement s'intéresser aux propriétés asymptotiques des graphes aléatoires mais d'introduire un objet limite et d'étudier les propriétés de celui-ci.

La théorie de la limite de graphes denses (i.e. quand le degré moyen croît à la même vitesse que le nombre de sommets) est apparu en 2006, initiée par les travaux de Lovász et Szegedy [139] et Borgs, Chayes, Lovász, Sós et Vesztergombi [38]. L'existence et les premières propriétés de l'objet limite ont été établies en 2006 par Lovasz et Szegedy [139]. L'objet limite, appelé graphon, peut s'exprimer comme une fonction symétrique et mesurable de deux variables. De manière informelle, un graphon peut être vu comme la version continue de la matrice d'adjacence d'un graphe dont l'ensemble des sommets est l'ensemble continu $[0, 1]$. L'unicité de la limite à une bijection préservant la mesure près a été prouvée par Borgs, Chayes et Lovász [36] en 2010. La première notion de convergence développée par Lovász et Szegedy [139] et Borgs, Chayes, Lovász, Sós et Vesztergombi [38] était une convergence locale définie en termes de densités d'homomorphismes (appelée aussi convergence à gauche). Une suite de graphes denses sera dite convergente si toutes les densités d'homomorphismes convergent simultanément pour tous les graphes simples finis (en particulier, les densités d'arêtes, les densités de triangles ...). C'est cette convergence qui nous intéressera dans la suite de la thèse.

D'autres notions de convergence ont été développées par la suite. Ainsi, une convergence métrique a été définie par Borgs, Chayes, Lovász, Sós et Vesztergombi [38, 40] à partir de la distance de coupe introduite par Frieze et Kannan [96] en 1999 qui permet de comparer des graphes de taille différente. Une suite de graphes denses est dite convergente pour la distance

de coupe, si elle est de Cauchy pour cette distance. Il a été prouvé que pour une suite de graphes denses, la convergence des densités homomorphismes est équivalente à la convergence pour la distance de coupe, voir Borgs, Chayes, Lovász, Sós et Vesztergombi [40, 41]. D'autres notions de convergence globale ont été introduites, voir Borgs, Chayes et Gamarnik [34] et Borgs, Chayes, Lovász, Sós et Vesztergombi [41]. Toutes ces convergences sont équivalentes dans le cas des suites de graphes denses.

Il a également été établi que tout graphon W est limite d'une suite de graphes denses, voir Lovász et Szegedy [139]. Une telle suite de graphes est donnée par les W-graphes aléatoires (« W-random graphs » en anglais), construits à partir du graphon W par échantillonnage.

La théorie des limites de suites de graphes creux, qui ne fera pas l'objet de notre étude, est également très riche mais plus difficile à appréhender. Il n'y a plus de notion de convergence métrique et la notion de convergence en termes de densités d'homomorphismes n'est plus pertinente puisque toute suite de graphes creux tend vers le graphon trivial nul.

La première notion de convergence naturelle pour les suites de graphes creux dont tous les sommets sont de degré borné est une convergence locale. Elle a été développée par Benjamini et Schramm [19] au début des années 2000. En fait, il a été montré que cette convergence est équivalente à la convergence des sous-graphes, définie par Bollobás et Riordan [33].

D'autres notions de convergence globales pour les suites de graphes creux, plus fortes que la convergence locale, ont également été proposées, voir Bollobás et Riordan [33], Borgs, Chayes et Gamarnik [34], Hatami, Lovász et Szegedy [103], Borgs, Chayes, Kahn et Lovász [35] et Lovász [140]. Contrairement aux limites de graphes denses, les liens entre les différentes notions de convergence ne sont pas encore totalement établis.

La théorie des limites de graphes est en pleine évolution et beaucoup de problèmes restent encore à explorer. Le livre de Lovász [138] publié en 2012 constitue un ouvrage de référence sur le sujet.

Résumé des travaux

Le premier article, écrit avec Jean-François Delmas et Jean-Stéphane Dhersin,

Cost functionals for large (uniform and simply generated) random trees, [54]

considère des fonctionnelles de coût sur les arbres aléatoires et plus précisément les arbres binaires sous le modèle de Catalan et les arbres simplement générés. Nous établissons des principes d'invariance ainsi que les fluctuations associées. Nos résultats de convergence reposent sur les limites d'échelle d'arbres discrets conditionnés à être grands. Dans le cas des arbres binaires sous le modèle de Catalan, on construit un arbre binaire comme un sous-arbre de l'arbre continu brownien codé par l'excursion brownienne normalisée. Ce plongement permet d'obtenir des convergences presque sûres des fonctionnelles additives, étendant ainsi la convergence en loi donnée par Fill et Kapur [89] et Fill et Janson [88]. Pour le modèle des arbres simplement générés, on utilise la convergence du processus de contour des arbres de Galton-Watson conditionnés vers le processus hauteur d'un arbre de Lévy stable. L'article est repris sans modification dans le Chapitre 3.

Le second article, écrit avec Jean-François Delmas et Jean-Stéphane Dhersin,

Asymptotic for the cumulative distribution function of the degrees and homomorphism densities for random graphs from a graphon, [53]

s'intéresse au comportement asymptotique de la fonction de répartition des degrés et des densités d'homomorphismes pour de grands graphes aléatoires échantillonnés à partir de graphons.

Nous obtenons la convergence des lois finies-dimensionnelles de la fonction de répartition empirique des degrés vers un processus gaussien centré dont nous donnons le noyau de covariance. Nous étendons ainsi les travaux de Bickel, Chen et Levina [25]. En remplaçant la fonction indicatrice par des fonctions plus régulières, nous obtenons la convergence presque sûre de suites de mesures construites à partir de densités d'homomorphismes de graphes partiellement étiquetés. La preuve de ce résultat est très différente de celle sur la fonction de répartition empirique des degrés. Nous mettons également en évidence les fluctuations associées à cette convergence. Ces résultats généralisent les travaux de Féray, Méliot and Nikeghbali [84].

L'article est reporté sans modification dans le Chapitre 4.

Organisation de la thèse

Ce manuscrit comporte une première partie introductive qui se décompose en deux chapitres indépendants :

- Le chapitre 1 constitue une introduction au chapitre 3 concernant les fonctionnelles additives sur les arbres aléatoires : les différents modèles d'arbres aléatoires et les principaux résultats sur les limites d'échelle d'arbres de Galton-Watson conditionnés sont rappelés. Nous donnons ensuite de nombreux exemples de fonctionnelles additives dans des domaines aussi variés que l'informatique ou la biologie. Après un rappel sur les résultats existants de convergence de fonctionnelles additives, nous énonçons les résultats principaux de l'article [54].

- Le chapitre 2 est une introduction au chapitre 4 consacré aux asymptotiques de fonctionnelles sur les graphes aléatoires échantillonnés à partir d'un graphon : la théorie des graphons ainsi que les propriétés principales y sont développées. Nous rappelons les résultats de convergence associés aux W-graphes aléatoires générés à partir d'un graphon W : la convergence presque-sûre des densités d'homomorphismes ainsi que les fluctuations associées. Nous concluons ce chapitre par l'énoncé de nos résultats présentés dans l'article [53].

La partie II constitue le cœur de cette thèse. Les chapitres 3 et 4 correspondent aux deux travaux présentés précédemment. Ils y sont reportés sans modification. Chaque chapitre de cette thèse peut globalement être lu indépendamment des autres.

Partie I

Présentation des résultats

CHAPITRE 1

FONCTIONNELLES ADDITIVES SUR LES ARBRES ALÉATOIRES

1.1 Arbres discrets

Un arbre discret est un graphe sans cycle, non orienté et connexe (on peut toujours relier deux nœuds de l'arbre par un chemin). Dans la suite, nous allons considérer des arbres enracinés et ordonnés. Un arbre est dit enraciné si un nœud est distingué comme étant la racine notée $\emptyset$. Il est alors possible de décrire un arbre par sa suite des générations : la racine est la génération 0, les voisins de la racine forment la génération 1 et plus généralement, les nœuds à distance k de la racine forment la k-ième génération. Si un nœud u de la génération k a des voisins dans la génération $k+1$ alors ces voisins sont appelés les enfants de u. Un arbre est dit ordonné si les enfants de chaque nœud sont munis d'un ordre. Autrement dit, les enfants du nœud u peuvent être ordonnés dans une suite $u_1, \ldots, u_d$ où d est le nombre d'enfants de u. Les arbres enracinés et ordonnés sont aussi parfois appelés arbres planaires.

1.1.1 Notions sur les arbres discrets

Nous allons définir les arbres planaires par leur ensemble de nœuds en utilisant le formalisme de Neveu [153]. On désigne par $\mathcal{U} = \bigcup_{n\geq 0} (\mathbb{N}^*)^n$ l'ensemble des suites finies d'entiers strictement positifs. Par convention, $(\mathbb{N}^*)^0 = \{\emptyset\}$. Pour $n \geq 0$ et $u \in (\mathbb{N}^*)^n \subset \mathcal{U}$, la longueur de u est notée $|u| = n$. Si $u = (u_1, u_2, \ldots, u_n) \in \mathcal{U}\backslash\emptyset$, alors on appelle $\breve{u} = (u_1, u_2, \ldots, u_{n-1})$ le parent de u (u est alors un enfant de $\breve{u}$). Soient $v, w \in \mathcal{U}$. La concaténation des suites v et w est notée vw avec la convention que $v\emptyset = v$ et $\emptyset w = w$. On dit que v est un ancêtre de u (au sens large) et on écrit $v \preccurlyeq u$ s'il existe $w \in \mathcal{U}$ tel que $u = vw$. Si $v \preccurlyeq u$ et $v \neq u$, alors on écrit $v \prec u$. Ainsi, $\preccurlyeq$ désigne l'ordre généalogique sur $\mathcal{U}$. On peut également définir l'ordre lexicographique sur $\mathcal{U}$ noté $<$. Par exemple, $\emptyset < 1 < (1,2) < (2,1) < (2,2)$. Remarquons que $u \prec v$ implique que $u < v$ mais la réciproque est fausse ($(1,2) < (1,3)$ mais $(1,2) \not\prec (1,3)$).

Définition. *Un arbre planaire* $\mathbf{t}$ *est un sous-ensemble de* $\mathcal{U}$ *satisfaisant les trois propriétés suivantes :*

- *Un arbre contient toujours la racine :* $\emptyset \in \mathbf{t}$,
- *Si un arbre contient un individu* $u \in \mathbf{t}\backslash\{\emptyset\}$ *alors il contient son parent* $\breve{u}$ *: si* $u \in \mathbf{t}\backslash\{\emptyset\}$, *alors* $\breve{u} \in \mathbf{t}$.
- *Chaque individu a un nombre fini d'enfants : pour tout* $u \in \mathbf{t}$, *il existe* $k_u(\mathbf{t}) \in \mathbb{N}$ *tel que, pour chaque* $i \in \mathbb{N}^*$, $ui \in \mathbf{t}$ *si et seulement si* $1 \leq i \leq k_u(\mathbf{t})$.

Pour $u \in \mathbf{t}$, l'entier $k_u(\mathbf{t})$ représente le nombre de descendants du nœud u. Le nœud u est appelée feuille (resp. nœud interne) si $k_u(\mathbf{t}) = 0$ (resp. $k_u(\mathbf{t}) > 0$). Le nœud $\emptyset$ est appelé la racine de $\mathbf{t}$. On note $|\mathbf{t}| = \text{Card}\ (\mathbf{t})$ le nombre de nœuds de $\mathbf{t}$ et on dit que $\mathbf{t}$ est fini si $|\mathbf{t}| < +\infty$. On note $\mathbb{T}$ l'ensemble des arbres planaires, $\mathbb{T}_0$ l'ensemble des arbres planaires finis et $\mathbb{T}^{(p)} = \{\mathbf{t} \in \mathbb{T},\ |\mathbf{t}| = p\}$ l'ensemble des arbres planaires à p nœuds. Rappelons que le nombre d'arbres planaires à p nœuds est donné par le $p-1$-ième nombre de Catalan (voir Drmota [68], partie 1.2.2) :

$$C_{p-1} = \frac{1}{p}\binom{2p-2}{p-1} = \frac{(2p-2)!}{(p-1)!\,p!}.$$

La hauteur H d'un arbre $\mathbf{t} \in \mathbb{T}$ est définie par :

$$H(\mathbf{t}) = \sup\{|u| : u \in \mathbf{t}\}.$$

On définit également le sous-arbre $\mathbf{t}_u \in \mathbb{T}$ de $\mathbf{t}$ « au dessus » de u par :

$$\mathbf{t}_u = \{v \in \mathcal{U},\ uv \in \mathbf{t}\}.$$

Ces sous-arbres sont appelés « fringe-trees » en anglais, ils ont été introduits par Aldous [7] (voir aussi Holmgren et Janson [109]). La figure 1.1 donne un exemple de sous-arbre.

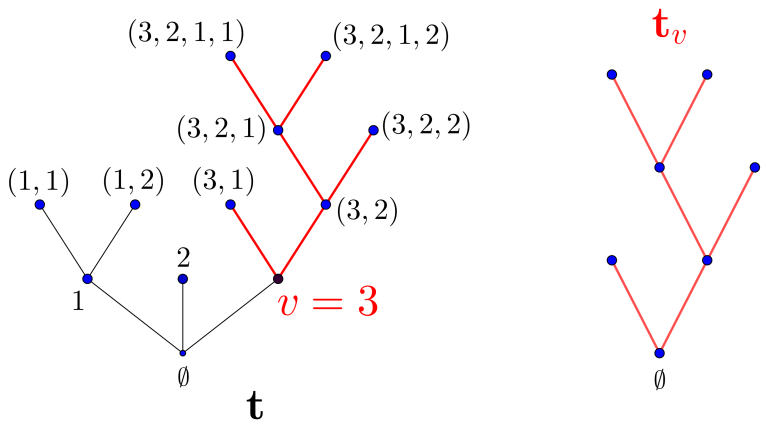


FIGURE 1.1 – Un arbre planaire à 5 nœuds internes et son sous-arbre issu du nœud $v = 3$ dessiné en rouge à droite de la figure.

On peut coder un arbre planaire fini par des fonctions, voir figure 1.2 (voir par exemple Le Gall [135]). Commençons par la fonction de contour (ou chemin de Dick). Pour cela, rappelons en quoi consiste le parcours d'un arbre en profondeur. Il s'agit de parcourir l'arbre, en partant de la racine $\emptyset$ au temps $t = 0$, et, arrivé en un nœud u, d'aller vers le premier enfant non visité ui ou alors, si tous les enfants ont déjà été visités, de retourner vers son parent. Le procédé s'arrête lorsqu'on est revenu à la racine en ayant visité tous les enfants de celle-ci. Étant donné que chaque arête est visitée deux fois, il est clair que le temps total d'exploration de l'arbre est $2(|\mathbf{t}| - 1)$. Pour $\mathbf{t} \in \mathbb{T}$, la fonction de contour de $\mathbf{t}$ est la fonction $C^{\mathbf{t}} : [0, 2(|\mathbf{t}| - 1)] \to \mathbb{R}^+$ qui associe à chaque étape $k \in [\![0, 2(|\mathbf{t}| - 1)]\!]$ du parcours en profondeur, la génération de l'individu v_k visité. On interpole ensuite linéairement $C^{\mathbf{t}}$ sur l'intervalle $[0, 2(|\mathbf{t}| - 1)]$, et l'on pose $C^{\mathbf{t}}(t) = 0$ pour tout $t > 2(|\mathbf{t}| - 1)$.

On peut également coder un arbre par sa fonction de hauteur. La fonction de hauteur de $\mathbf{t} \in \mathbb{T}$ est la suite de générations des individus de $\mathbf{t}$, quand ces individus sont classés dans l'ordre lexicographique. Si $\mathbf{t} \in \mathbb{T}$, on énumère les sommets dans l'ordre lexicographique $u_0 = \emptyset < u_1 < \cdots < u_{|\mathbf{t}|-1}$. Pour chaque $n \in [\![0, |\mathbf{t}| - 1]\!]$, on définit $h_{\mathbf{t}}(n) = |u_n|$ comme la hauteur du sommet u_n. On pose $h_{\mathbf{t}}(m) = 0$ pour $m \geq |\mathbf{t}|$, et on prolonge alors $h_{\mathbf{t}}$ à $\mathbb{R}^+$ par

interpolation linéaire. La fonction $h_{\mathbf{t}} = (h_{\mathbf{t}}(t), t \geq 0)$ est appelée fonction de hauteur de $\mathbf{t}$. La fonction de hauteur est similaire à la fonction de contour mais chaque individu de l'arbre n'est visité qu'une fois.

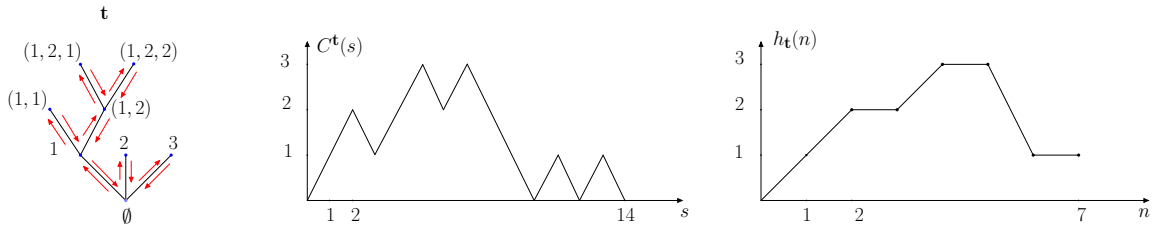


Figure 1.2 – Un arbre $\mathbf{t}$, son processus de contour $C^{\mathbf{t}}$ et sa fonction de hauteur $h_{\mathbf{t}}$

Dans la suite nous allons nous intéresser à deux types d'arbres aléatoires : les arbres aléatoires binaires (auxquels on associe généralement deux modèles classiques, le modèle de Catalan et le modèle de permutations aléatoires) et les arbres simplement générés (et plus particulièrement les arbres de Galton-Watson conditionnés par leur taille totale).

Nous commençons par quelques rappels sur les arbres binaires. Le lecteur pourra se référer au livre de Drmota [68], parties 1.2.1 et 1.4.1, Mahmoud [142], chapitre 2 et Sedgewick et Flajolet [177], chapitre 6. Ces arbres ont de nombreuses applications notamment en informatique au travers des algorithmes de tri. Ces algorithmes ont été étudiés à la fin des années 1960 et au début des années 1970 par Knuth [126, 128], qui donne des analyses détaillées de plusieurs paramètres combinatoires qui permettent de déterminer par exemple leur performance moyenne. Sedgewick [176], quant à lui, s'est intéressé à un des modèles de tri les plus utilisés, l'algorithme Quicksort, qui a été inventé par Hoare au début des années 1960.

1.1.2 Arbres binaires

Définition

Un arbre binaire complet est un arbre enraciné et ordonné où chaque nœud a zéro (c'est une feuille) ou deux enfants (c'est un nœud interne). Un arbre binaire à n nœuds internes a $n+1$ nœuds externes et donc $2n+1$ nœuds au total. On note $\mathbb{T}_{\text{bin}}^{(n)}$ l'ensemble des arbres binaires complets enracinés à n nœuds internes.

Arbres binaires sous le modèle de Catalan

Le plus simple pour choisir au hasard un arbre dans un ensemble d'arbres est le choix uniforme. Ce modèle est aussi appelé modèle de Catalan. Il sera l'objet de notre étude dans le chapitre 3. Soit T_n une variable aléatoire à valeurs dans $\mathbb{T}_{\text{bin}}^{(n)}$ choisie uniformément dans l'ensemble fini $\mathbb{T}_{\text{bin}}^{(n)}$. Comme le nombre d'arbres binaires complets avec n nœuds internes est donné par le n-ième nombre de Catalan $C_n = \frac{1}{n+1}\binom{2n}{n}$, on obtient que la probabilité d'obtenir un arbre particulier est l'inverse du n-ième nombre de Catalan. La loi de T_n est donnée par

$$\mathbb{P}(\mathrm{T}_n = \mathbf{t}) = \frac{1}{C_n}, \text{ pour tout } \mathbf{t} \in \mathbb{T}_{\text{bin}}^{(n)}.$$

Le choix uniforme d'un arbre parmi un ensemble d'arbres peut également s'appliquer à d'autres types d'arbres que les arbres binaires. Par analogie, on parlera encore de « modèle de Catalan » à ne pas confondre avec ce qu'on appelle les arbres de Catalan qui sont des arbres planaires.

Arbres marqués

Dans la suite, il conviendra de faire la distinction entre deux types d'arbres : les arbres non marqués et les arbres marqués. Dans le premier cas, les nœuds des arbres ne contiennent pas d'information et l'on s'intéresse donc simplement à leur forme. Dans le second cas, les arbres sont vus comme des structures de données c'est-à-dire que les nœuds contiennent des informations aussi appelées clés ou marques. Si l'on efface les clés d'un arbre marqué, l'arbre non marqué obtenu est appelé sa forme. Formellement, un arbre marqué est une paire $\tilde{\mathbf{t}} = (\mathbf{t}, (\gamma_v, v \in \mathbf{t}))$ où les marques γ_v sont des éléments d'un ensemble Γ. L'arbre non marqué $\mathbf{t}$ est appelé le squelette ou la forme de l'arbre marqué $\tilde{\mathbf{t}}$, voir la définition donnée par Neveu [153]. C'est l'ensemble des nœuds de l'arbre marqué.

Arbres binaires de recherche et modèle de permutations uniformes

Une des applications les plus importantes des arbres binaires est l'algorithme des arbres binaires de recherche (ABR en abrégé ou encore BST pour « binary search tree » en anglais), voir Mahmoud [142], chapitre 2 ou Chauvin, Clément et Gardy[47], partie 1.2.5. Les ABR sont des structures de données fondamentales en informatique. Ils sont utiles pour classer et chercher des données, appelées des clés. Un ABR à n nœuds internes est un arbre binaire marqué par des clés $x_1, \dots, x_n$ prises dans un ensemble totalement ordonné Γ qui satisfait la contrainte que la clé de chaque nœud interne est plus grande que toutes les clés de son sous-arbre de gauche et plus petite que toutes les clés de son sous-arbre de droite. On peut rajouter $n+1$ feuilles non étiquetées à cette structure afin d'obtenir un arbre binaire complet marqué à n nœuds internes. Le squelette d'un ABR à n nœuds internes est donc un arbre binaire complet à n nœuds internes et donc $n+1$ feuilles et $2n+1$ nœuds.

Une généralisation de ces arbres sont les arbres m-aire de recherche (en anglais, m-ary search trees), voir Mahmoud [142], partie 1.7, Drmota [68], partie 1.4.2 ou Holmgren et Janson [109], où chaque nœud peut contenir jusqu'à $m-1$ éléments d'un ensemble totalement ordonné.

On peut munir l'ensemble des ABR à n nœuds internes d'une loi de probabilité. Le modèle le plus étudié dans la littérature est le modèle de permutations uniformes (RPM en anglais pour « random permutation model »), voir Fill [85], Mahmoud [142], chapitre 2 ou Chauvin, Clément et Gardy[47], partie 2.2.1. Un ABR à n nœuds internes peut être généré par une permutation (ou liste) π de l'ensemble $[\![1, n]\!]$. A chaque permutation de $[\![1, n]\!]$ est associé un ABR. Dans le modèle de permutations uniformes, chaque permutation π de l'ensemble $[\![1, n]\!]$ est équiprobable. Remarquons que des permutations différentes peuvent engendrées le même arbre squelette (voir figure 1.3). Ainsi, contrairement au modèle de Catalan, tous les arbres squelettes ne sont pas équiprobables. Soit $\tilde{\mathrm{T}}_n$ une variable aléatoire à valeurs dans l'ensemble des arbres binaires complets marqués choisie selon le modèle de permutations uniformes. On note T_n l'arbre squelette associé à $\tilde{\mathrm{T}}_n$ qui est une variable aléatoire à valeurs dans $\mathbb{T}_{\mathrm{bin}}^{(n)}$. La loi de T_n est donnée par

$$Q(\mathbf{t}) = \mathbb{P}\left(\mathrm{T}_n = \mathbf{t}\right) = \frac{1}{\prod_{v \in \mathbf{t}} |\mathbf{t}_v|}, \text{ pour tout } \mathbf{t} \in \mathbb{T}_{\mathrm{bin}}^{(n)}. \tag{1.1}$$

L'ABR aléatoire issu d'une permutation uniforme de taille n a une forme très différente d'un arbre binaire de Catalan : sa hauteur (resp. sa longueur totale, ou longueur de cheminement, égale à la somme des distances des différents nœuds à la racine) est asymptotiquement d'ordre $\log n$ (resp. d'ordre $n \log n$) alors qu'elle est asymptotiquement d'ordre $\sqrt{n}$ (resp. d'ordre $n\sqrt{n}$) sous le modèle de Catalan. Ainsi, le modèle de permutations uniformes donne plus de poids aux arbres dits équilibrés que le modèle de Catalan (voir figure 1.3).

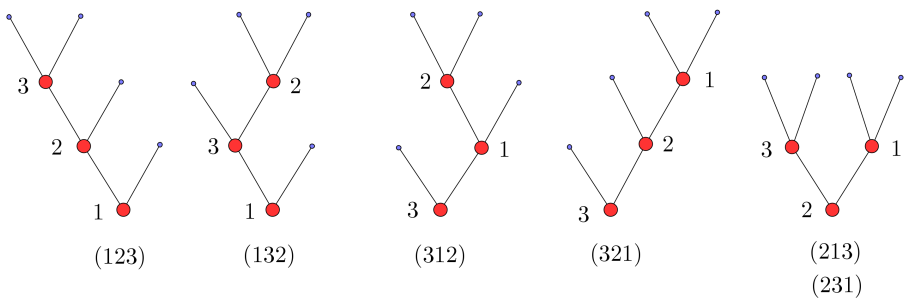


Figure 1.3 – $C_3 = 5$ arbres binaires complets ordonnés et enracinés à 3 nœuds internes

Faisons maintenant quelques rappels concernant les arbres de Galton-Watson qui constituent une classe importante d'exemples d'arbres aléatoires ordonnés et enracinés.

1.2 Arbres de Galton-Watson

1.2.1 Définitions

Soit $\mathfrak{p} = (\mathfrak{p}(n), n \in \mathbb{N})$ une loi de probabilité sur $\mathbb{N}$. Le processus de Galton-Watson $Z = (Z_n, n \in \mathbb{N})$ de loi de reproduction $\mathfrak{p}$ décrit l'évolution de la taille d'une population issue d'un individu et dont les individus ont chacun un nombre d'enfants suivant la loi $\mathfrak{p}$, de façon indépendante les uns des autres. Autrement dit, si on introduit une famille $(\xi_{i,n}; i \in \mathbb{N}, n \in \mathbb{N})$ de variables aléatoires indépendantes de loi $\mathfrak{p}$, on définit le processus Z par récurrence par

$$Z_0 = 1 \quad \text{et } Z_{n+1} = \sum_{i=1}^{Z_n} \xi_{i,n+1}.$$

L'arbre de Galton-Watson associé au processus de Galton-Watson Z représente l'arbre généalogique de la population où chaque individu est relié à son parent et à ses descendants. Z_n représente le nombre de nœuds au niveau k dans l'arbre de Galton-Watson et $\sum_{k\geq 0} Z_k$ le nombre de nœuds de l'arbre de Galton-Watson :

Définition. *Soit $\mathfrak{p}$ une mesure de probabilité sur $\mathbb{N}$. Soit τ une variable aléatoire à valeurs dans $\mathbb{T}$. On dit que τ est un arbre de Galton-Watson de loi de reproduction $\mathfrak{p}$ si l'on a :*

- $\mathbb{P}(k_\emptyset(\tau) = j) = \mathfrak{p}(j)$ *pour tout* $j \in \mathbb{N}$,
- *pour tout $j \in \mathbb{N}^*$, sous $\mathbb{P}(\tau \,|\, k_\emptyset(\tau) = j)$, les j sous-arbres $\tau_1, \ldots, \tau_j$ de τ issus de $\emptyset$ sont indépendants et de même loi que τ.*

On peut montrer que cette loi existe et est uniquement caractérisée par $\mathfrak{p}$. En fait, τ est un arbre de Galton-Watson de loi de reproduction $\mathfrak{p}$ si et seulement si pour tout $h \in \mathbb{N}^*$ et $\mathbf{t} \in \mathbb{T}$ tel que $H(\mathbf{t}) \leq h$, on a :

$$\mathbb{P}\left(r_h(\tau) = \mathbf{t}\right) = \prod_{\substack{u \in \mathbf{t} \\ |u| < h}} \mathfrak{p}(k_u(\mathbf{t})),$$

où pour tout $\mathbf{t} \in \mathbb{T}$ et pour tout $h \in \mathbb{N}^*$, $r_h(\mathbf{t})$ est le sous-arbre de $\mathbf{t}$ obtenu en gardant tous les nœuds à hauteur inférieure ou égale à h et est défini par : $r_h(\mathbf{t}) = \{u \in \mathbf{t} : |u| \leq h\}$. En particulier, quand $\tau \in \mathbb{T}_0$, on a la formule suivante :

$$\mathbb{P}\left(\tau = \mathbf{t}\right) = \prod_{u \in \mathbf{t}} \mathfrak{p}(k_u(\mathbf{t})), \text{ pour tout } \mathbf{t} \in \mathbb{T}_0. \tag{1.2}$$

La probabilité $\mathbb{P}(\tau = \mathbf{t})$ est ce qu'on appellera le poids de $\mathbf{t}$ dans le cadre des arbres simplement générés.

On exclut le cas trivial où $\mathfrak{p}(1) = 1$. On rappelle que le processus de Galton-Watson Z est dit sous-critique (resp. critique, sur-critique) si le nombre moyen d'enfants $m = \sum_{k\in\mathbb{N}} k\mathfrak{p}(k)$ satisfait $m < 1$ (resp. $m = 1$, $m > 1$). Un arbre de Galton-Watson de loi de reproduction $\mathfrak{p}$ est fini presque-sûrement si $m \leq 1$ et infini avec probabilité strictement positive si $m > 1$ (voir Harris [102], chapitre 1 ou Athreya et Ney [15], chapitre 3).

1.2.2 Arbres de Galton-Watson conditionnés et arbres simplement générés

Dans cette partie, nous allons nous intéresser aux arbres de Galton-Watson conditionnés à avoir un certain nombre de nœuds. Nos résultats asymptotiques sur les fonctionnelles additives du Chapitre 3 concerneront en particulier ce modèle. En fait, les arbres de Galton-Watson conditionnés rentrent dans le cadre plus général des arbres simplement générés. Nous préciserons les liens qui existent entre ces deux modèles d'arbres. Nous renvoyons à Drmota [68], partie 1.2.7 et Janson [114], chapitre 2 pour les définitions et des propriétés des arbres de Galton-Watson conditionnés et des arbres simplement générés.

Les arbres de Galton-Watson ont des tailles aléatoires. On peut cependant s'intéresser à des arbres aléatoires avec des tailles fixées. Soit $\mathfrak{p}$ une loi de reproduction. Pour tout entier n suffisamment grand tel que $\mathbb{P}(|\tau| = n) > 0$, on définit alors $\tilde{\tau}^{(n)}$ comme un arbre de Galton-Watson conditionné à avoir n nœuds, c'est à dire un arbre aléatoire de loi :

$$\mathbb{P}(\tilde{\tau}^{(n)} = \mathbf{t}) = \frac{\mathbb{P}(\tau = \mathbf{t})}{\mathbb{P}(|\tau| = n)}, \text{ pour tout } \mathbf{t} \in \mathbb{T} \text{ tel que } |\mathbf{t}| = n.$$

Par définition, la taille de $\tilde{\tau}^{(n)}$ est égale à n.

Nous allons maintenant définir les arbres simplement générés qui sont une généralisation des arbres de Galton-Watson conditionnés par leur taille totale. Ils ont été introduits par Meir et Moon [144] en 1978. On considère une suite de poids $\mathfrak{q} = (\mathfrak{q}(k), k \in \mathbb{N})$ de réels positifs telle que $\sum_{k\in\mathbb{N}} \mathfrak{q}(k) > \mathfrak{q}(1) + \mathfrak{q}(0)$ et $\mathfrak{q}(0) > 0$ (autrement dit, $\mathfrak{q}(0) > 0$ et il existe $k \geq 2$ tel que $\mathfrak{q}(k) > 0$). Ces hypothèses sont faites pour éviter les cas triviaux. Pour tout $\mathbf{t} \in \mathbb{T}_0$, on définit le poids $w(\mathbf{t})$ de $\mathbf{t}$ par :

$$w(\mathbf{t}) = \prod_{v\in\mathbf{t}} \mathfrak{q}(k_v(\mathbf{t})).$$

On pose pour $p \in \mathbb{N}^*$:

$$w(\mathbb{T}^{(p)}) = \sum_{\mathbf{t}\in\mathbb{T}^{(p)}} w(\mathbf{t}).$$

Définition. *Pour $p \in \mathbb{N}^*$ tel que $w(\mathbb{T}^{(p)}) > 0$, un arbre simplement généré à valeurs dans $\mathbb{T}^{(p)}$ et associé à la suite de poids $\mathfrak{q}$ est une variable aléatoire $\tau^{(p)}$ à valeurs dans $\mathbb{T}^{(p)}$ dont la loi est caractérisée par :*

$$\mathbb{P}(\tau^{(p)} = \mathbf{t}) = \frac{w(\mathbf{t})}{w(\mathbb{T}^{(p)})}, \quad \mathbf{t} \in \mathbb{T}^{(p)}.$$

En particulier, d'après (1.2), si la suite de poids $\mathfrak{q}$ est une loi de probabilité sur $\mathbb{N}$ alors $\tau^{(p)}$ a la même loi qu'un arbre de Galton-Watson de loi de reproduction $\mathfrak{q}$ et conditionné à avoir p nœuds.

Même quand $\mathfrak{q}$ n'est pas une loi de probabilité, on peut sous certaines conditions sur la loi $\mathfrak{q}$, voir les arbres simplement générés comme des arbres de Galton-Watson conditionnés, voir Janson [114], section 4. On dit que deux suites de poids $\mathfrak{q} = (\mathfrak{q}(k), k \in \mathbb{N})$ et $\tilde{\mathfrak{q}} = (\tilde{\mathfrak{q}}(k), k \in \mathbb{N})$ sont équivalentes s'ils existent $a, b > 0$ tels que :

$$\tilde{\mathfrak{q}}(k) = ab^k \mathfrak{q}(k), \quad \forall k \in \mathbb{N}.$$

Il est facile de voir que deux suites de poids équivalentes définissent les mêmes arbres simplement générés. On introduit la fonction génératrice $g_{\mathfrak{q}}$ de $\mathfrak{q}$: $g_{\mathfrak{q}}(\theta) = \sum_{k\in\mathbb{N}} \theta^k \mathfrak{q}(k)$ pour $\theta > 0$. Notons $\rho \in [0, +\infty]$ son rayon de convergence. Dans [114], section 4, Janson montre qu'il existe une loi de probabilité équivalente à la suite de poids $\mathfrak{q}$ si et seulement si $\rho > 0$. Dans ce cas, les lois de probabilité équivalentes à $\mathfrak{q}$ sont données par :

$$\mathfrak{p}(k) = \frac{t^k\, \mathfrak{q}(k)}{g_{\mathfrak{q}}(t)}, \qquad \text{pour tout } t > 0 \text{ tel que } g_{\mathfrak{q}}(t) < \infty.$$

Ainsi, un arbre simplement généré dont la suite de poids $\mathfrak{q}$ vérifie la condition précédente peut-être défini par une loi de probabilité $\mathfrak{p}$ équivalente à $\mathfrak{q}$ et donc être vu comme un arbre de Galton-Watson conditionné. Ceci signifie que $\tilde{\tau}^{(p)}$ associée à $\mathfrak{p}$ et $\tau^{(p)}$ associé à $\mathfrak{q}$ ont même loi. Il est également possible dans certains cas de se ramener à une loi de probabilité critique. On dit que $\mathfrak{q}$ est générique si l'équation $sg'_{\mathfrak{q}}(s) = g_{\mathfrak{q}}(s)$ admet une unique solution $s_{\mathfrak{q}}$. Dans ce cas, la loi de probabilité $\mathfrak{p}$ définie par

$$\mathfrak{p}(k) = \frac{s_{\mathfrak{q}}^k\, \mathfrak{q}(k)}{g_{\mathfrak{q}}(s_{\mathfrak{q}})}, \quad k \geq 0,$$

est une suite de poids équivalente à $\mathfrak{q}$. De plus, la probabilité $\mathfrak{p}$ est une loi de probabilité critique, voir Janson [114], section 4.

Plusieurs classes combinatoires d'arbres aléatoires à n nœuds peuvent être réalisées comme des arbres simplement générés à n nœuds et donc aussi comme des arbres de Galton-Watson conditionnés à avoir n nœuds, voir table 1.1. C'est le cas des arbres planaires à n nœuds, des arbres de Cayley (arbres enracinés, étiquetés et non ordonnés) à n nœuds et des arbres binaires complets à n nœuds à condition que n soit impair (et donc a $\frac{n-1}{2}$ nœuds internes), que l'on munit de la probabilité uniforme (i.e. dans chaque classe, chaque arbre est équiprobable), voir Aldous [9], Drmota [68], partie 1.2.7 ou Janson [114], chapitre 10.

Une des propriétés remarquables des arbres simplement générés est qu'ils ont tendance à avoir une structure déséquilibrée, i.e. ils contiennent quelques branches de grandes tailles. Ainsi, la hauteur moyenne d'un arbre simplement généré de taille n est d'ordre $\sqrt{n}$ (par opposition à un arbre de taille n dit équilibré qui aura une hauteur d'ordre $\log(n)$) (voir [68], chapitres 3 et 4 ou Janson [114], section 21.3).

Remarquons ici que nous nous sommes intéressés aux arbres de Galton-Watson conditionnés par leur taille totale, mais d'autres conditionnements ont également été étudiés comme par exemple, avoir une grande hauteur ou avoir un grand nombre de feuilles. Par opposition aux limites d'échelle des arbres de Galton-Watson, qui est une convergence globale, on peut ainsi s'intéresser aux limites locales de ces arbres. Dans son papier fondateur [122] de 1986, Kesten a été le premier à s'intéresser aux limites locales d'un arbre de Galton-Watson critique ou sous-critique conditionné à avoir une grande hauteur. Il a montré que la limite est l'arbre biaisé par la taille avec une unique branche infinie. Pour d'autres références sur les limites locales d'arbres de Galton-Watson conditionnés avec des conditionnements plus généraux, voir Jonsson et Stefánsson [119], Janson [114], chapitre 7 ou encore Abraham et Delmas [4, 3].

1.2.3 Convergence du processus de contour des arbres de Galton-Watson conditionnés

On commence par définir l'excursion brownienne normalisée notée $B_{\text{ex}} = (B_{\text{ex}}(t), 0 \leq t \leq 1)$. De manière informelle, B_{ex} est un mouvement brownien standard partant de l'origine et conditionné à rester strictement positif sur $(0, 1)$ et à revenir en 0 au temps 1. Il existe plusieurs manières de définir et caractériser une excursion brownienne (voir Revuz et Yor [168], chapitre 12, partie 2 ou McKean [112], partie 2.9). Biane [23], Verwaat [189] ou encore

Classe d'arbres à n nœuds munie de la probabilité uniforme	Arbres planaires	Arbres de Cayley	Arbres binaires complets
Cardinal de l'ensemble	C_{n-1}	n^{n-1}	$C_{\frac{n-1}{2}}$
Suite de poids de l'arbre simplement généré	$\mathfrak{p}(k) = 1,\ \forall k \geq 0$	$\mathfrak{p}(k) = \frac{1}{k!},\ \forall k \geq 0$	$\mathfrak{p}(0) = \mathfrak{p}(2) = 1$ et $\mathfrak{p}(k) = 0,\ \forall k \neq 0, 2$
Fonction génératrice associée à la suite de poids	$g_{\mathfrak{p}}(t) = \frac{1}{1-t}$	$g_{\mathfrak{p}}(t) = e^t$	$g_{\mathfrak{p}}(t) = 1 + t^2$
Loi critique équivalente π	Loi géométrique de paramètre $\frac{1}{2}$	Loi de Poisson de paramètre 1	$\pi(0) = \frac{1}{2}$ et $\pi(2) = \frac{1}{2}$
Variance σ^2 de π	$\sigma^2 = 2$	$\sigma^2 = 1$	$\sigma^2 = 1$

TABLE 1.1 – Exemples d'arbres aléatoires vus comme des arbres de Galton-Watson conditionnés.

Rogers et Williams [171], partie 40 ont montré que l'excursion brownienne normalisée peut être construite à partir d'un pont brownien (processus stochastique dont la loi est celle d'un mouvement brownien standard (i.e. issu de 0) et conditionné à valoir 0 au temps 1), lui-même pouvant être décrit en fonction d'un mouvement brownien standard. En fait, l'excursion brownienne normalisée a même loi qu'un pont de Bessel de dimension 3 i.e. que la norme d'un pont brownien de dimension 3. Ainsi, pour représenter une trajectoire d'une excursion brownienne normalisée, on construit un pont brownien de dimension 3 noté $X(t) = (X_1(t), X_2(t), X_3(t))$ à partir d'un mouvement brownien de dimension 3, $B(t) = (B_1(t), B_2(t), B_3(t))$ de la manière suivante :

$$X_i(t) = B_i(t) - tB_i(1), \text{ pour tout } t \in [0,1],\ i \in \{1,2,3\}.$$

Enfin, l'excursion brownienne normalisée B_{ex} est obtenue de la manière suivante :

$$B_{\text{ex}}(t) = \|X(t)\| = \sqrt{X_1(t)^2 + X_2(t)^2 + X_3(t)^2}, \text{ pour tout } t \in [0,1].$$

On note $\mathbb{N}^{(1)}$ la loi de l'excursion brownienne normalisée.

Soit $\mathfrak{p}$ une loi de reproduction sur $\mathbb{N}$ critique (i.e. $\sum_{k\in\mathbb{N}} k\mathfrak{p}(k) = 1$). On note $\sigma^2 > 0$ sa variance finie ou infinie. Soit τ un arbre de Galton-Watson de loi de reproduction $\mathfrak{p}$ et pour chaque p tel que $\mathbb{P}(|\tau| = p) > 0$, soit $\tau^{(p)}$ un arbre de même loi que τ conditionnellement à $\{|\tau| = p\}$. Nous allons donner les résultats de convergence du processus de contour de $\tau^{(p)}$.

La variance de $\mathfrak{p}$ est finie : Dans le cas où la variance σ^2 est finie, Aldous [10] a montré que le processus de contour $C^{\tau^{(p)}}$ de $\tau^{(p)}$ correctement normalisé converge en loi vers l'excursion brownienne normalisée.

Theorem (Aldous). *On suppose que la loi de reproduction* $\mathfrak{p}$ *est critique et que* $0 < \sigma^2 < +\infty$. *Alors on a la convergence en loi dans l'espace* $\mathcal{C}([0,1])$ *muni de la convergence uniforme :*

$$\frac{\sigma}{2\sqrt{p}}\left(C^{\tau^{(p)}}(2ps), s \in [0,1]\right) \xrightarrow[p\to+\infty]{(d)} (B_{ex}(s), s \in [0,1]),$$

où la convergence a lieu pour la sous-suite infinie de p *tels que* $\mathbb{P}(|\tau| = p) > 0$.

La loi $\mathfrak{p}$ est dans le domaine d'attraction d'une loi stable : Une généralisation du théorème d'Aldous dans la cas où $\mathfrak{p}$ appartient au domaine d'attraction d'une loi stable a été obtenue par Duquesne [69]. On suppose donc que $\mathfrak{p}$ est dans le domaine d'attraction d'une loi stable. Rappelons la définition. Soit $(U_k, k \in \mathbb{N}^*)$ une suite de variables indépendantes de même loi de distribution $\mathfrak{p}$. On pose $W_p = \sum_{k=1}^{p} U_k - p$. On dit que $\mathfrak{p}$ est dans le domaine d'attraction d'une loi stable d'exposant de Laplace $\psi(\lambda) = \kappa\lambda^\gamma$ avec $\gamma \in (1,2]$ et $\kappa > 0$, et de suite de normalisation $(a_p, p \in \mathbb{N}^*)$ de réels positifs ou nuls, si W_p/a_p converge en loi, quand p tend vers l'infini, vers une variable aléatoire X d'exposant de Laplace ψ (i.e. on a $\mathbb{E}[\mathrm{e}^{-\lambda X}] = \mathrm{e}^{-\psi(\lambda)}$ pour $\lambda \geq 0$).

Remarque. Si $\mathfrak{p}$ a une variance σ^2 finie alors on peut choisir $a_p = \sqrt{p}$ et X est alors de loi normale centrée de variance σ^2, de telle sorte que $\psi(\lambda) = \sigma^2\lambda^2/2$.

Theorem (Duquesne). *On suppose que la loi de reproduction* $\mathfrak{p}$ *est critique et qu'elle est dans le domaine d'attraction d'une loi stable d'exposant* $\psi(\lambda) = \kappa\lambda^\gamma$ *avec* $\gamma \in (1,2]$ *et* $\kappa > 0$, *et de suite de normalisation* $(a_p, p \in \mathbb{N}^*)$. *Alors il existe un processus aléatoire non trivial, continu, positif ou nul* $H = (H_s, s \in [0,1])$, *tel que on ait la convergence en loi dans l'espace* $\mathcal{C}([0,1])$ *muni de la convergence uniforme :*

$$\frac{a_p}{p}\left(C^{\tau^{(p)}}(2ps), s \in [0,1]\right) \xrightarrow[p\to+\infty]{(d)} H,$$

où la convergence a lieu pour la sous-suite infinie de p *tels que* $\mathbb{P}(|\tau| = p) > 0$.

Le processus H, voir Duquesne [69] pour une construction de H, désigne une excursion normalisée du processus hauteur, introduit dans Duquesne et Le Gall [70], d'un arbre de Lévy de mécanisme de branchement ψ.

1.3 Arbres réels

Dans cette partie, on introduit la notion d'arbres réels, définis en tant qu'espaces métriques. Nous verrons que ces arbres peuvent être codés par une fonction continue de la même manière que les arbres discrets peuvent l'être avec leur fonction de contour.

1.3.1 Définitions

Nous allons définir les arbres réels ou $\mathbb{R}$-arbres en terme d'espaces métriques (voir Evans [78]) :

Définition. *Un arbre réel (resp. arbre réel compact) est un espace métrique (resp. espace métrique compact)* $(\mathcal{T}, d)$ *qui satisfait les deux propriétés suivantes pour tout* $x, y \in \mathcal{T}$ *:*

- *Il existe une unique isométrie* $f_{x,y}$ *de* $[0, d(x,y)]$ *dans* $\mathcal{T}$ *telle que l'on ait* $f_{x,y}(0) = x$ *et* $f_{x,y}(d(x,y)) = y$.
- *Si* ϕ *est une application continue injective de* $[0,1]$ *dans* $\mathcal{T}$ *telle que* $\phi(0) = x$ *et* $\phi(1) = y$, *alors on a* $\phi([0,1]) = f_{x,y}([0, d(x,y)])$.

Un arbre réel enraciné est un arbre réel $(\mathcal{T}, d)$ ayant un point distingué $\emptyset$ appelé la racine.

De manière équivalente, un espace métrique $(\mathcal{T}, d)$ est un arbre réel si et seulement si :

- $\mathcal{T}$ est connecté
- d satisfait la condition suivante appelée « condition des 4 points » :

$$d(s,t) + d(x,y) \leq \max(d(s,x) + d(t,y), d(s,y) + d(t,x)) \quad \text{pour tout} \quad s,t,x,y \in \mathcal{T}.$$

On peut munir l'espace des arbres réels enracinés et compacts d'une distance appelée distance de Gromov-Hausdorff qui rend cet espace polonais, voir par exemple Le Gall [135], théorème 2.1.

De la même manière que la fonction de contour code les planaires, on peut coder certains arbres réels par une fonction continue. Soit h une fonction continue de $[0,1]$ dans $\mathbb{R}^+$ telle que $h(0) = h(1) = 0$. Pour $x, y \geq 0$. On définit

$$d_h(x,y) = h(x) + h(y) - 2 \min_{z \in [x \wedge y, x \vee y]} h(z),$$

où $x \wedge y = \min(x,y)$ et $x \vee y = \max(x,y)$. Il est facile de vérifier que d_h est symétrique et satisfait l'inégalité triangulaire et donc que c'est une pseudo-distance sur $[0,1]$. On introduit une relation d'équivalence $\sim_h$ définie sur $[0,1]^2$ par $x \sim_h y \Leftrightarrow d_h(x,y) = 0$. L'arbre continu codé par h est l'espace métrique quotient $\mathcal{T}_h = [0,1]/\sim_h$ muni de la distance d_h et enraciné en la classe d'équivalence de 0. Il est facile de voir que $(\mathcal{T}_h, d_h)$ est un arbre réel compact (voir Le Gall et Miermont [137], Théorème 3.1). On note $\mathbf{p}_h$ la projection canonique de $[0,1]$ dans $\mathcal{T}_h$. Pour tout $t \in [0,1]$, $\mathbf{p}_h(t)$ est donc un nœud de $\mathcal{T}_h$ à distance $h(t)$ de la racine. Pour $0 \leq s \leq t \leq 1$, l'ancêtre commun aux nœuds $\mathbf{p}_h(s)$ et $\mathbf{p}_h(t)$ est à distance $\min_{s \leq r \leq t} h(s)$ de la racine. La figure 1.4 représente le sous-arbre de l'arbre $\mathcal{T}_h$ engendré par les feuilles $\mathbf{p}_h(s), \mathbf{p}_h(t), \mathbf{p}_h(u)$ et la racine.

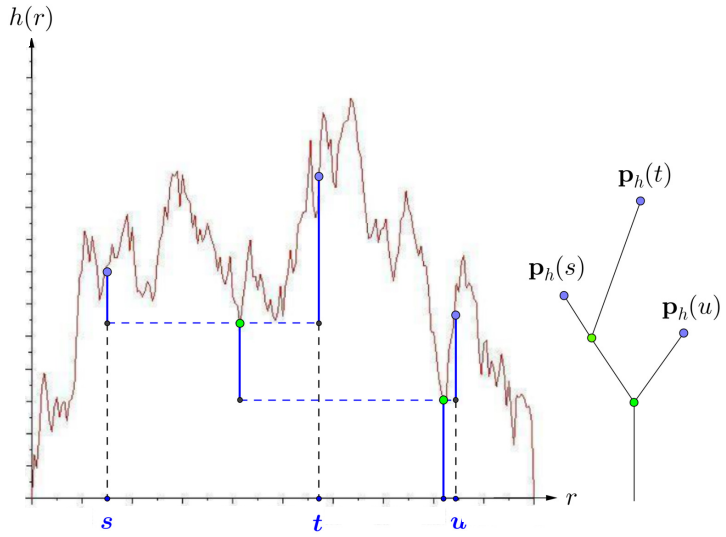


FIGURE 1.4 – Un sous-arbre de l'arbre $\mathcal{T}_h$

En fait, il est possible de montrer que tout arbre réel compact peut être codé par une fonction continue (voir Le Gall [135], partie 2.2).

1.3.2 Arbre brownien

Dans une série de papiers [8, 9, 10] datant du début des années 1990, Aldous introduit et développe la notion d'arbre continu brownien (d'abord appelé arbre continu compact) ainsi que plusieurs résultats de convergence. On rappelle que $B_{\text{ex}} = (B_{\text{ex}}(t) : 0 \leq t \leq 1)$ désigne l'excursion brownienne normalisée.

Définition. *L'arbre continu brownien (abrégé en CRT pour « continuum random tree » en anglais) est l'arbre aléatoire réel $\mathcal{T}_{B_{ex}}$ codé par l'excursion brownienne normalisée B_{ex}.*

Cette définition a été donnée par Aldous dans [10], corollaire 22 avec une différence d'un facteur 2 i.e. le CRT est codé par $2B_{\text{ex}}$ et non B_{ex}. Ainsi, afin d'énoncer nos résultats dans les deux cas précédents, il sera utile d'introduire le processus $e = \sqrt{2/\alpha} B_{\text{ex}}$ avec $\alpha > 0$. En effet, pour $\alpha = 1/2$, on est dans le cadre de l'arbre brownien défini par Aldous et pour $\alpha = 2$, on est dans le cadre de l'arbre brownien associé à l'excursion brownienne normalisée B_{ex}.

Comme e est un processus stochastique, $\mathcal{T}_e$ est une variable aléatoire dans l'espace polonais des arbres aléatoires réels enracinés et compacts. De plus, puisque les minima locaux du mouvement brownien (et donc de l'excursion brownienne e) sont toujours distincts presque sûrement, $\mathcal{T}_e$ est un arbre binaire presque sûrement (i.e. les nœuds ont 0 ou 2 enfants).

Dans [8, 9, 10], Aldous énonce quatre autres définitions équivalentes pour construire l'arbre brownien réel :

- par séparation poissonnienne d'une droite (« Poisson line-breaking » en anglais), voir Aldous [8] ou Pitman [163], partie 7.4.
- comme limite d'échelle d'arbres de Galton-Watson ayant une loi limite critique et de variance finie, conditionnés à avoir un grand nombre de nœuds, voir Aldous [9] et Le Gall [135].
- comme limite d'échelle d'arbres non étiquetés et non enracinés au sens de Gromov-Hausdroff (conjecture d'Aldous [9] p55, et prouvé en 2014 par Stufler [181]).
- par ses marginales finies dimensionnelles i.e. la loi de ses sous-arbres engendrés/couverts (« spanning trees » en anglais) par des feuilles choisies uniformément au hasard (voir Aldous [10], partie 4.3). Cette dernière construction sera développée dans la section suivante.

1.3.3 Construction d'arbres binaires à partir de l'arbre continu brownien

Dans cette partie, on considère des arbres marqués $\tilde{\mathbf{t}} = (\mathbf{t}, (h_v, v \in \mathbf{t}))$ où $h_v \geq 0$, pour tout $v \in \mathbf{t}$. h_v est la longueur de la branche en dessous de v et $\mathbf{t}$ est le squelette de l'arbre marqué $\tilde{\mathbf{t}}$, voir partie 1.1.2. On note $\tilde{\mathbb{T}}^{(p)}$ désigne l'ensemble des arbres marqués à p feuilles. Si $\tilde{\mathbf{t}} = (\mathbf{t}, (h_v, v \in \mathbf{t}))$ est un arbre marqué, la longueur de $\tilde{\mathbf{t}}$ est définie par :

$$L(\tilde{\mathbf{t}}) = \sum_{v \in \mathbf{t}} h_v.$$

Soient $n \in \mathbb{N}$ et $0 < t_1 < \cdots < t_{n+1} < 1$. On considère le sous-arbre de l'arbre continu brownien τ_e engendré par les $n+1$ feuilles $\mathbf{p}_e(t_1), \ldots, \mathbf{p}_e(t_{n+1})$ et la racine $\emptyset$. On le note $\tau_e(t_1, \ldots, t_{n+1})$. C'est un arbre binaire, enraciné, complet et ordonné à $n+1$ feuilles et $2n+1$ arêtes (n'oublions pas que la racine possède également une arête, voir figure 1.4). On peut lui associer un arbre marqué

$$\tilde{\mathbf{t}}(e; t_1, \ldots, t_{n+1}) = (\mathbf{t}, \{h_v(e; t_1, \ldots, t_{n+1}), v \in \mathbf{t}\}),$$

où intuitivement, $\mathbf{t} = \mathbf{t}(e; t_1, \ldots, t_{n+1})$ est identique à $\tau_e(t_1, \ldots, t_{n+1})$ mais avec des longueurs de branches égalent à 1 et n'ayant pas de branche sous la racine. Pour tout $v \in \mathbf{t}$,

$h_v(e;t_1,\dots,t_{n+1})$ est la longueur de la branche dans $\tau_e(t_1,\dots,t_{n+1})$ sous le nœud correspondant à $v \in \mathbf{t}$. Ainsi, l'arbre squelette $\mathbf{t}$ est un arbre discret complet, ordonné et enraciné à $n+1$ feuilles et $2n$ arêtes. Pour une construction plus rigoureuse, voir Aldous [10], Le Gall [134, 135] ou encore Duquesne et Le Gall [70].

Un processus stochastique peut être décrit par ses lois finies dimensionnelles. Inspirés par cette idée, nous allons caractériser l'arbre continu brownien en terme de ses marginales finies dimensionnelles qui sont ses sous-arbres couverts par des feuilles choisies uniformément au hasard.

On définit la mesure uniforme Λ_n sur l'ensemble des arbres marqués dont le squelette appartient à l'ensemble $\mathbb{T}_{\text{bin}}^{(n)}$ des arbres binaires complets et ordonnés à n nœuds internes (et donc $n+1$ feuilles) par :

$$\int \Lambda_n(d\tilde{\mathbf{t}})F(\tilde{\mathbf{t}}) = \sum_{\mathbf{t}\in\mathbb{T}_{(n)}^{\text{bin}}} \int \prod_{v\in\mathbf{t}} dh_v \, F(\mathbf{t},\{h_v, v\in\mathbf{t}\}),$$

pour toute fonction F positive, bornée et continue sur l'ensemble des arbres marqués à n feuilles.

Le théorème suivant a d'abord été énoncé par Aldous [9], partie 2.4. On peut également trouver l'énoncé de ce résultat dans Le Gall [134] et [135] partie 2.6 et Pitman [163], chapitre 7. On rappelle que $\mathbb{N}^{(1)}$ est la loi de l'excursion brownienne normalisée.

Theorem (Aldous)**.** *La loi de l'arbre $\tilde{\mathbf{t}} = \tilde{\mathbf{t}}(e;t_1,\dots,t_{n+1})$ sous la mesure de probabilité*

$$(n+1)!\mathbf{1}_{\{0\le t_1\le\dots\le t_{n+1}\le 1\}} dt_1\dots dt_p \mathbb{N}^{(1)}(de)$$

est

$$(n+1)!\, 2\,\alpha^{n+1} L(\tilde{\mathbf{t}}) \exp(-2L(\tilde{\mathbf{t}})^2)\Lambda_n(d\tilde{\mathbf{t}}).$$

Autrement dit, pour toute fonction F positive bornée et continue sur l'ensemble des arbres marqués à $n+1$ feuilles, on a :

$$\int \mathbb{N}^{(1)}(de) \int_{0\le t_1\le\dots\le t_{n+1}\le 1} F\left(\tilde{\mathbf{t}}(e;t_1,\dots,t_{n+1})\right) = 2\alpha^{n+1} \int \Lambda_n(d\tilde{\mathbf{t}}) L(\tilde{\mathbf{t}}) \exp(-2L(\tilde{\mathbf{t}})^2) F(\tilde{\mathbf{t}}).$$

De ce théorème, il est facile de voir que le squelette $\mathbf{t}(e;t_1,\dots,t_{n+1})$ de $\tilde{\mathbf{t}}(e;t_1,\dots,t_{n+1})$ est uniformément distribué sur l'ensemble des arbres binaires complets ordonnés à $n+1$ feuilles dont le cardinal est donné par le n-ième nombre de Catalan : $C_n = (2n)!/(n!(n+1)!)$. Ainsi, conditionnellement au squelette, sous la mesure de probabilité

$$(n+1)!\mathbf{1}_{\{0\le t_1\le\dots\le t_{n+1}\le 1\}} dt_1\dots dt_p \mathbb{N}^{(1)}(de),$$

la densité du vecteur $(h_v(e;t_1,\dots,t_{n+1}), v\in\mathbf{t})$ est donnée par :

$$C_n 2\,\alpha^{n+1}(n+1)!\, L(\tilde{\mathbf{t}}) \exp(-2L(\tilde{\mathbf{t}})^2) = 2\frac{(2n)!}{n!}\alpha^{n+1} L(\tilde{\mathbf{t}}) \exp(-2L(\tilde{\mathbf{t}})^2).$$

Les longueurs d'arêtes sont des variables échangeables et sont indépendantes du squelette.

On se donne maintenant $(U_n, n\in\mathbb{N}^*)$ une suite de variable aléatoires indépendantes de loi uniforme sur $[0,1]$ et indépendante de e. On note $\mathcal{T}_{[n]}$ l'arbre aléatoire réel engendré par les $n+1$ feuilles $\mathbf{p}(U_1),\dots,\mathbf{p}(U_{n+1})$ et la racine et $\tilde{\mathrm{T}}_n = (\mathrm{T}_n;(h_{n,v}, v\in\mathrm{T}_n))$ l'arbre marqué

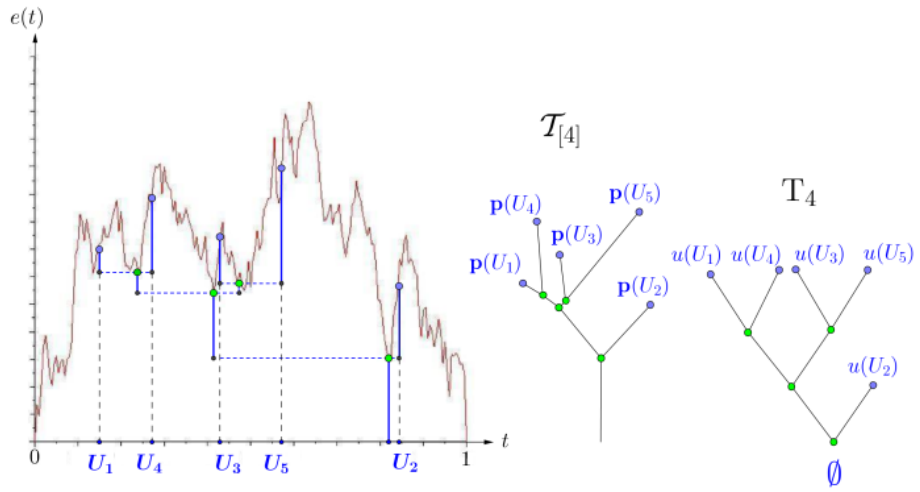


Figure 1.5 – L'excursion brownienne, $\tilde{\mathrm{T}}_n$ et T_n (pour $n = 4$).

associé (voir figure 1.5 pour une simulation avec $n = 4$). On note L_n la longueur de l'arbre marqué $\tilde{\mathrm{T}}_n$.

Le théorème précédent se réécrit ainsi de la manière suivante : la loi de probabilité de $\tilde{\mathrm{T}}_n$ est

$$(n+1)!\,2\,\alpha^{n+1} L_n \exp(-2L_n^2)\Lambda_{n+1}(d\tilde{\mathrm{T}}_n),$$

et la densité de $(h_{n,v}, v \in \mathrm{T}_n)$ est, conditionnellement à T_n, donnée par :

$$f_n((h_{n,v}, v \in \mathrm{T}_n)) = 2\frac{(2n)!}{n!}\,\alpha^{n+1}\,L_n\,\mathrm{e}^{-\alpha L_n^2} \prod_{v\in \mathrm{T}_n} \mathbf{1}_{\{h_{n,v}>0\}}. \tag{1.3}$$

En particulier, L_n est indépendante de T_n et a même loi que $\sqrt{\Delta_n/\alpha}$ où Δ_n est une loi Gamma de paramètre $n+1$ et 1 (voir Aldous [10] ou Pitman [163], Théorème 7.9). Ainsi, conditionnellement à T_n, la densité de L_n est donnée par :

$$f_{L_n}(x) = 2\frac{\alpha^{n+1}}{n!} x^{2n+1}\,\mathrm{e}^{-\alpha x^2}\,\mathbf{1}_{\{x>0\}}. \tag{1.4}$$

1.4 Fonctionnelles de coût et fonctions péage

Les fonctionnelles de coût sur les arbres planaires ont été largement étudiées dans des domaines très variés (informatique, physique, chimie, biologie ...), voir les références dans les parties 1.4.2 et 1.4.3. Nous allons étudier l'asymptotique de ces fonctionnelles définies sur les arbres planaires (arbres enracinés et ordonnés) quand le nombre de nœuds tend vers l'infini. Les fonctionnelles que nous allons considérer sont additives (définies par une relation de récurrence) et induites par une fonction dite péage. La valeur d'une fonctionnelle additive sur un arbre est définie de manière récursive comme la somme de la valeur de la fonctionnelle sur les sous-arbres enracinés sur les enfants de la racine de l'arbre à laquelle on ajoute un terme (le péage) qui est une fonction qui dépend de la taille de l'arbre. Les fonctionnelles de coût définies de la sorte représentent le coût des algorithmes du type diviser et conquérir, où la nature récursive intrinsèque de ces algorithmes amène naturellement à une telle formulation.

1.4.1 Définition

Définition. *Une fonctionnelle de coût F sur les arbres planaires est une fonctionnelle additive sur les arbres si elle satisfait la relation de récurrence suivante :*

$$F(\mathbf{t}) = \sum_{i=1}^{k_\emptyset(\mathbf{t})} F(\mathbf{t}_i) + b_{|\mathbf{t}|}$$

pour tout arbre $\mathbf{t} \in \mathbb{T}_0$ tel que $|\mathbf{t}| \geq 1$ et avec $F(\emptyset) = 0$ et où $\mathbf{t}_1, \ldots, \mathbf{t}_{k_\emptyset(\mathbf{t})}$ sont les sous-arbres enracinés aux enfants de la racine de $\mathbf{t}$. La suite $(b_k, k \geq 1)$ est appelée la fonction péage.

Par définition de F, il est facile de voir que l'on a pour tout $\mathbf{t} \in \mathbb{T}_0$:

$$F(\mathbf{t}) = \sum_{v \in \mathbf{t}} b_{|\mathbf{t}_v|}. \tag{1.5}$$

Remarque.

1. En particulier, une fonctionnelle de coût F sur les arbres binaires est une fonctionnelle additive sur les arbres si elle satisfait la relation de récurrence suivante :

   $$F(\mathbf{t}) = F(L(\mathbf{t})) + F(R(\mathbf{t})) + b_{|\mathbf{t}|}$$

   pour tout arbre $\mathbf{t}$ tel que $|\mathbf{t}| \geq 1$ et avec $F(\emptyset) = 0$ et où $L(\mathbf{t}) = \mathbf{t}_1$ (resp. $R(\mathbf{t}) = \mathbf{t}_2$) désigne le sous-arbre de gauche (resp. de droite) de la racine de $\mathbf{t}$.

2. La fonction péage peut également dépendre de manière plus générale de l'arbre lui même. Nos résultats concerneront le cas où elle ne dépend que du cardinal de l'arbre, voir Flajolet, Gourdon et Martínez [92], Devroye [57] et Wagner [190].

1.4.2 Transition de phase

Par la suite, on distinguera les fonctions de coût dites locales et globales. Dans le chapitre 3, nous nous restreindrons à l'étude de fonctionnelles globales. Ces deux termes sont utilisés pour mettre en évidence la transition de phase entre le régime normal et non-normal des lois limites. Intuitivement, lorsque l'on considère des fonctions de coût dites locales, i.e. des fonctions de coût petites, la somme est dominée par la contribution des très nombreux petits sous-arbres. Comme les différentes parties des arbres ont une très faible dépendance les unes entres les autres, cela rend possible la normalité asymptotique. Pour une fonction péage dite globale, i.e. une fonction qui croit très rapidement avec la taille de l'arbre, la somme sera au contraire dominée par la contribution des très grands arbres, qui sont plus fortement dépendants les uns des autres, faisant donc apparaître d'autres distributions limites.

Quand la fonction de coût dépend seulement de la taille de l'arbre, la transition de phase entre les différents régimes pour les arbres binaires a été étudiée par Hwang et Neininger [111] pour le modèle de permutations uniformes et par Fill et Kapur [89] pour le modèle de Catalan.

Dans le cas où les fonctions de coût dépendent de l'arbre lui-même Wagner [191] étudie la transition de phase pour des arbres simplement générés, des arbres récursifs et des arbres binaires de recherche. Le fait que les fonctions de coût peuvent dépendre de l'arbre lui-même amplifie les relations de dépendance rendant difficile la détermination de la zone de transition de phase.

Nous allons consacrer la fin de ce paragraphe aux fonctionnelles locales, qui bien que n'entrant pas dans notre cadre d'étude, ont été très étudiées. Nous en donnons quelques exemples : taille totale ($b_k = 1$), nombre de feuilles ($b_k = \mathbf{1}_{\{k=1\}}$), nombre de nœuds protégés (un nœud est dit protégé si ce n'est pas une feuille ni le parent d'une feuille), nombre de sous-arbres de taille donnée p ($b_k = \mathbf{1}_{\{k=p\}}$) et nombre total de sous-arbres, voir Janson [115], partie 2.

On peut également considérer des fonctions péage qui dépendent plus généralement de l'arbre $\mathbf{t} \in \mathbb{T}$ et non plus simplement de sa taille $|\mathbf{t}|$: par exemple, on peut compter le nombre de sous-arbres précisément égaux à un sous-arbre donné $\mathbf{t}_0$ (« tree patterns » en anglais) dont la fonction péage est donnée par $\mathbf{1}_{\{\mathbf{t}=\mathbf{t}_0\}}$, voir Flajolet, Gourdon et Martínez [92], Devroye [57] et Wagner [190].

La fonctionnelle de forme (« shape functional » en anglais) $Q(\mathbf{t})$, définie dans (1.1), est également une fonctionnelle qui a suscité un grand intérêt. Elle a d'abord été étudiée dans le cadre des arbres binaires de recherche sous le modèle de permutations uniformes, voir Dobrow et Fill [64] ou Fill [85]. Elle est appelée fonctionnelle de forme car elle permet d'évaluer la forme de l'arbre $\mathbf{t}$ puisque les valeurs maximales de $Q(\mathbf{t})$ sont atteintes pour les arbres qui sont les plus équilibrés. La fonctionnelle additive $-\log Q$ est associée à la fonction péage $b_k = \log(k)$. Sa convergence a été étudiée par Fill [85, 89] pour les arbres binaires sous le modèle de Catalan et de permutations uniformes et Meir et Moon [145] pour les arbres simplement générés. Cette fonctionnelle intervient aussi dans une méthode d'auto-organisation des arbres binaires de recherche appelée « move-to-root », voir Dobrow et Fill [62, 63].

De nombreux articles fournissent des résultats asymptotiques pour des fonctionnelles additives locales assez générales. Ces études permettent notamment de retrouver le comportement limite des fonctionnelles citées précédemment. Le lecteur pourra se référer aux articles suivants : pour les arbres binaires de recherche sous le modèle de permutations uniformes, voir Devroye [60], Holmgren et Janson [108] et Wagner [191], pour les arbres de Galton-Watson conditionnées (i.e. les arbres simplement générés), voir Janson [115] et Wagner [191], pour les arbres récursifs, voir Holmgren et Janson [108] et Wagner [191], pour les arbres étiquetés, voir Wagner [190], pour les arbres m-aire de recherche, voir Fill et Kapur [90], et enfin pour les arbres m-aire croissants de recherche, voir Ralaivaosaona et Wagner [165].

1.4.3 Fonctionnelles de coût globales

Nous donnons maintenant différents exemples de fonctionnelles additives globales que l'on peut trouver dans la littérature et qui interviennent naturellement en informatique, physique ou biologie. Nos résultats plus généraux permettent d'obtenir des asymptotiques sur des fonctionnelles additives globales plus générales.

On note d la distance usuelle de graphe sur $\mathbf{t} \in \mathbb{T}$. Ainsi, pour $v, w \in \mathbf{t}$, w est un ancêtre de v (que l'on écrit $w \preccurlyeq v$, voir partie 1.1.1) si $d(\emptyset, v) = d(\emptyset, w) + d(w, v)$. Pour $u, v \in \mathbf{t}$, on note $u \wedge v$, l'ancêtre le plus récent de u et v : $u \wedge v$ est l'unique élément de $\mathbf{t}$ tel que : $w \preccurlyeq u$ et $w \preccurlyeq v$ implique que $w \preccurlyeq u \wedge v$.

Pour $\mathbf{t} \in \mathbb{T}$, on a par exemple les fonctionnelles suivantes :

- **Longueur de cheminement total** :

$$P(\mathbf{t}) = \sum_{v \in \mathbf{t}} d(\emptyset, v) = \sum_{v \in \mathbf{t}} |\mathbf{t}_v| - |\mathbf{t}|.$$

  La fonction péage est donnée par : $b_k = k - 1$. Dans les cas des arbres binaires de recherche, la fonctionnelle compte le nombre de comparaisons de l'algorithme de tri rapide (« quicksort » en anglais) pour trier une liste de nombre distincts, voir Rösler [172].

  Pour les arbres binaires sous le modèle de Catalan et les arbres de Galton-Watson conditionnés à avoir un certain nombre de nœuds, il est connu que cette fonctionnelle converge vers la loi de Airy, $2\int_0^1 B_s\, ds$ où B est un mouvement brownien standard sur

[0, 1], voir Takàcs [183], Aldous [9, 10] et Janson [113]. L'asymptotique de cette fonctionnelle a également été étudiée pour les arbres binaires sous le modèle de permutations uniformes, voir Régnier [166], Rösler [172] et Fill et Janson [87] et pour les arbres m-aire de recherche, voir Fill et Kapur [90, 91] .

- **Indice de Wiener** :

$$W(\mathbf{t}) = \sum_{u,v\in\mathbf{t}} d(u,v) = 2|T| \sum_{w\in T} |T_w| - 2\sum_{w\in T} |T_w|^2.$$

La fonctionnelle fait intervenir les fonctions de péage $b_k = k$ et $b_k = k^2$. L'indice de Wiener a été introduit en 1947 par Harold Wiener [193] comme le nombre de chemins (« path number » en anglais). Il était initialement défini comme le nombre de liens entre chaque paire d'atomes dans une molécule acyclique. L'indice de Wiener joue un rôle important dans les propriétés physiques-chimiques des structures chimiques (point d'ébullition, défauts de cristaux ...), voir [101, 66, 185].

La convergence a été étudiée par Janson [113] et Chassaing et Janson [116] pour les arbres binaires sous le modèle de Catalan, Neininger [152] pour les arbres binaires de recherche et les arbres récursifs et Janson [113] pour les arbres de Galton-Watson conditionnés à avoir un certain nombre de nœuds.

- **Fonctionnelles de coût associées aux fonctions péages de type puissance** : $b_k = k^\beta$ avec $\beta > 0$.

Fill et Kapur [89] ont montré la convergence en loi de la fonctionnelle pour les arbres binaires sous le modèle de Catalan. La limite est caractérisée en termes de ses moments. Leurs preuves reposent sur des calculs combinatoires. Plus tard, Fill et Janson [88] conjecturent une expression de la limite qui s'exprime en fonction de l'excursion normalisée. Le cas des arbres binaires de recherche a été traité par Neininger [151] (avec $\beta \in \mathbb{R}$, $\beta > 1$) et les arbres m-aire par Fill et Kapur [90, 91].

De nombreuses fonctionnelles globales étudiées dans la littérature se réécrivent en termes de fonctionnelles associées aux fonctions péages $b_k = k^\beta$ avec $\beta > 0$.

Fonctionnelles de coût sur les arbres phylogénétiques

De nombreux exemples de fonctionnelles proviennent également de la biologie au travers des arbres phylogénétiques qui sont utiles pour classer des populations et représenter l'évolution des relations entre les espèces (voir Aldous [13], Ford [93]).

Un arbre phylogénétique de taille n (aussi appelé « cladogram » en anglais) est un arbre binaire complet enraciné à n feuilles étiquetées de 1 à n qui correspondent aux espèces et à $n-1$ nœuds internes qui correspondent à leurs ancêtres, voir Ford [93]. Les longueurs de branches ne seront pas prises en compte, si bien que l'on s'intéressera seulement à la topologie des arbres i.e. à leur forme.

On peut munir les arbres phylogénétiques de deux modèles de probabilité : le modèle de Yule-Harding ou modèle d'évolution neutre, défini par Yule [195] en 1924 (qui correspond exactement au modèle de permutations uniformes sur les ABR puisque un arbre phylogénétique a n feuilles peut être vu comme un ABR à $n-1$ nœuds, voir Aldous [11]) et le modèle uniforme (i.e. le modèle de Catalan en informatique).

Quand le nombre d'espèces augmente, il est intéressant de regarder la forme des arbres afin de mieux comprendre leur structure. Par exemple, une asymétrie dans un arbre peut traduire une adaptation meilleure d'une ou plusieurs espèces. Nous donnons trois exemples d'indices sur les arbres phylogénétiques qui rentrent dans le cadre des fonctionnelles globales (voir Shao et Sokal [179], Heard [104], Kirkpatrick et Slatkin [123], Mooers et Heard [149], Felsenstein

[82], chapitre 33 pour d'autres exemples d'indice). L'intérêt de ces indices est qu'ils dépendent seulement de la forme des arbres et en particulier, ils sont invariants par isomorphisme et par renumérotation des feuilles. Les indices de Colless et cophénétique ne concernent que les arbres binaires.

- **Indice de Sackin (= longueur du chemin extérieur)** :

$$S(\mathbf{t}) = \sum_{v \in \mathcal{L}(\mathbf{t})} d(\emptyset, v),$$

où $\mathcal{L}(\mathbf{t})$ désigne l'ensemble des feuilles de l'arbre $\mathbf{t}$. L'indice de Sackin a été introduit en 1972 par Sackin [175].

Sa convergence a été étudiée par Blum, François et Janson [27] pour les arbres binaires pour le modèle de Catalan et le modèle de permutations uniformes. L'indice de Sackin a également été étudié en informatique dans le cadre des arbres binaires de recherche sous le modèle de permutations uniformes (donc sous le modèle de Yule), voir Régnier [166], Rösler [172]. Dans ce cas, cet indice est plus connu sous le nom de longueur de chemin extérieur (« external path length » en anglais) qui correspond au nombre de comparaisons dans l'algorithme quicksort.

- **Indice de Colless** :

$$C(\mathbf{t}) = \sum_{v \in T} \left| |\mathcal{L}(L_v)| - |\mathcal{L}(R_v)| \right|,$$

où $|\mathcal{L}(L_v)|$ (resp. $|\mathcal{L}(R_v)|$) est le nombre de feuilles du sous-arbre de gauche (resp. de droite) au dessus du nœud v. L'indice de Colless a quant à lui été introduit un peu plus tard, en 1982 par Colless [50], paragraphe 2.2, pour mesurer l'asymétrie d'un arbre binaire.

Sa convergence a été étudiée par Blum, François et Janson [27] pour le modèle de Catalan et le modèle de permutations uniformes.

- **Indice cophénétique** :

$$Co(\mathbf{t}) = \sum_{u,v \in \mathcal{L}(\mathbf{t}), u \neq v} d(\emptyset, u \wedge v).$$

Cet indice a été considéré récemment, en 2013, par Mir, Rosseló et Rotger [146] et Cardona, Mir et Rosseló [42] qui ont étudié certaines de ses propriétés asymptotiques sous les deux modèles, telles que son espérance et sa variance asymptotiques.

1.5 Résultats sur les fonctionnelles globales du Chapitre 3

Dans ce paragraphe, nous allons synthétiser les résultats du chapitre 3 qui correspond à l'article [54] paru dans *Electronic Journal of Probability*.

Le but de ce travail a été dans un premier temps, de prouver la conjecture posée par Fill et Janson [88] et donc de s'intéresser à l'étude asymptotique de fonctionnelles additives associées à des fonction de péage du type $b_n = n^\beta$ pour $n \in \mathbb{N}^*$ et $\beta > 0$ pour des arbres binaires sous le modèle de Catalan. Nous avons en fait montrer un résultat plus fort permettant d'atteindre des fonctionnelles plus générales. La convergence des fonctionnelles associées aux fonctions de péage du type puissance sera un cas particulier d'application. Nous nous sommes ensuite intéressés à des résultats similaires mais pour les arbres simplement générés.

1.5.1 Résultats pour les arbres binaires sous le modèle de Catalan

Contexte

Soit $(\mathrm{T}_n : n \in \mathbb{N}^*)$ une suite d'arbres binaires obtenus comme sous-arbres de l'arbre continu brownien selon la procédure d'échantillonnage décrite dans la partie 1.3.3. On rappelle que T_n a même loi qu'un arbre binaire à n nœuds internes sous le modèle de Catalan (i.e qu'un arbre choisi uniformément dans l'ensemble des arbres binaires complets et ordonnés ayant n nœuds internes et donc $n+1$ feuilles). En particulier, T_n possède $2n+1$ nœuds. Pour $v \in \mathrm{T}_n$, on note $\mathrm{T}_{n,v} = (\mathrm{T}_n)_v$ le sous-arbre de T_n au dessus du nœud v.

Espaces fonctionnels

Soit I un intervalle de $\mathbb{R}$ de mesure de Lebesgue strictement positive. On note $\mathcal{B}(I)$ l'espace des fonctions mesurables de I dans $\mathbb{R}$. On note $\mathcal{C}(I)$ l'espace des fonctions continues de I dans $\mathbb{R}$. Pour $f \in \mathcal{B}(I)$, on note $\| f \|_\infty$ la norme infinie et par $\| f \|_{\text{esssup}}$ la norme supremum essentiel de $|f|$ sur I. Les deux supremum coïncident quand f est continue.

Des mesures aléatoires

Pour $n \in \mathbb{N}^*$, on définit la mesure aléatoire pondérée A_n sur $[0,1]$ définie pour tout $f \in \mathcal{B}([0,1])$ par :

$$A_n(f) = |\mathrm{T}_n|^{-3/2} \sum_{v \in \mathrm{T}_n} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right).$$

Pour $h \in \mathcal{C}_+([0,1])$, on définit la longueur de l'excursion de h au dessus du niveau $r \in \mathbb{R}+$ et passant par $s \in [0,1]$ par :

$$\sigma_{r,s}(h) = \int_0^1 dt\, \mathbf{1}_{\{m_h(s,t) \geq r\}},$$

où pour $s,t \in [0,1]$, $m_h(s,t) = \inf_{u \in [s \wedge t, s \vee t]} h(u)$, voir figure 1.6.

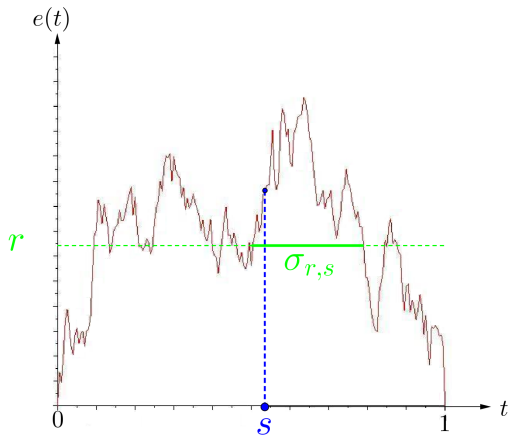


FIGURE 1.6 – Longueur d'une sous-excursion de e au dessus du niveau r et enjambant s.

Pour $h \in \mathcal{C}_+([0,1])$, on considère aussi la mesure Φ_h sur $[0,1]$ définie par :

$$\Phi_h(f) = \int_0^1 ds \int_0^{h(s)} dr\; f(\sigma_{r,s}(h)), \quad f \in \mathcal{B}([0,1]).$$

On munit l'espace des mesures finies positives sur $[0,1]$ de la topologie de la convergence étroite.

Une convergence presque sûre et ses fluctuations

On obtient la convergence presque sûre pour toutes les fonctions continues sur $[0,1]$ qui explosent de manière contrôlée en 0 :

Théorème A. *On a presque sûrement, pour tout $f \in \mathcal{B}([0,1])$, continue sur $(0,1]$ et telle qu'il existe $a \in [0,1/2)$ tel que $\lim_{x\to 0+} x^a f(x) = 0$:*

$$A_n(f) \underset{n\to\infty}{\longrightarrow} \sqrt{2\alpha}\,\Phi_e(f).$$

Remarque. La fonctionnelle de forme ne peut être atteinte par nos résultats : en effet, la fonction f définie par $f(x) = \log(x)/x$ ne satisfait pas les hypothèses du théorème A.

La preuve repose sur l'approximation de l'arbre brownien réel $\mathcal{T}$ par un sous-arbre binaire $\mathcal{T}_{[n]}$, voir la construction rappelée dans la partie 1.3.3, puis sur l'approximation dans l'arbre marqué associé $\tilde{\mathrm{T}}_n$, des longueurs de branche $(h_{n,v} : v \in \mathrm{T}_n)$ par leur valeur moyenne que l'on déduit de (1.3). La mesure aléatoire A_n est ainsi approchée par plusieurs mesures aléatoires et chacune des approximations successives est contrôlée en norme L^2.

On obtient également les fluctuations associées à cette convergence presque sûre :

Théorème B. *Soit $f \in \mathcal{C}([0,1])$ une fonction localement lipschitzienne continue sur $(0,1]$ telle qu'il existe $a \in (0,1)$ tel que $\| x^a f' \|_{esssup}$ soit finie. On a la convergence en loi suivante :*

$$\left(|\mathrm{T}_n|^{1/4}(A_n - \sqrt{2\alpha}\,\Phi_e)(f),\ A_n\right) \xrightarrow[n\to\infty]{(d)} \left((2\alpha)^{1/4}\sqrt{\Phi_e(xf^2)}\ G,\ \sqrt{2\alpha}\ \Phi_e\right),$$

où G est une variable aléatoire de loi normale centrée, réduite indépendante de l'excursion brownienne e.

On constate dans la preuve que les fluctuations proviennent de l'estimation des longueurs de branche $(h_{n,v} : v \in \mathrm{T}_n)$ par leur moyenne dans l'arbre marqué $\tilde{\mathrm{T}}_n$ et non pas de l'approximation de l'arbre continu brownien $\mathcal{T}$ par le sous-arbre $\mathcal{T}_{[n]}$.

Application aux fonctions de coût du type puissance

Dans toute cette partie, on considère le cas $\alpha = 2$ de sorte que e est l'excursion brownienne normalisée B_{ex}.

On applique les théorèmes A et B à la fonction f définie par $f(x) = x^{\beta-1}$, pour tout $x \in [0,1]$, avec $\beta > 0$. On pose $Z_\beta := \Phi_{B_{\mathrm{ex}}}(x^{\beta-1})$ et on définit une suite de variables aléatoires $(Z_\beta^{(n)} : n \in \mathbb{N}^*)$ par :

$$Z_\beta^{(n)} := A_n(x^{\beta-1}) = |\mathrm{T}_n|^{-(\beta+\frac{1}{2})} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta, \quad \forall n \in \mathbb{N}^*,$$

D'après les hypothèses du théorème A, on obtient directement la convergence presque sûre pour tout $\beta > \frac{1}{2}$. Pour $0 < \beta \leq \frac{1}{2}$, un argument de convergence monotone en β est nécessaire. Nous obtenons finalement la convergence presque sûre suivante :

Proposition C. *On a presque sûrement pour tout $\beta > 0$:*

$$\lim_{n\to+\infty} Z_\beta^{(n)} = 2Z_\beta.$$

On montre également que pour $\beta > 1$,

$$2Z_\beta = \beta \int_0^1 \left[t^{\beta-1} + (1-t)^{\beta-1}\right] B_{\mathrm{ex}}(t)dt - \beta(\beta-1)\int_{0<s<t<1} (t-s)^{\beta-2}\left[B_{\mathrm{ex}}(s) + B_{\mathrm{ex}}(t) - 2m_{B_{\mathrm{ex}}}(s,t)\right] dsdt, \quad (1.6)$$

et que $2Z_1 = 2\int_0^1 B_{\mathrm{ex}}(s)ds$. Ces formules correspondent également à celles données par Fill et Janson [88]. Elle se simplifient dans le cas où $\beta > 1$ à :

$$2Z_\beta = \beta(\beta-1)\int_{[0,1]^2} |t-s|^{\beta-2} m_{B_{\mathrm{ex}}}(s,t)dsdt.$$

Fill et Janson [88] conjecturent que l'égalité (1.6) est également vraie pour tout $1/2 < \beta < 1$. Pour $\beta > 1/2$, Fill et Kapur [89] ont montré que $Z_\beta^{(n)}$ converge en loi vers $2Z_\beta$ et ont obtenu tous les moments de Z_β pour $\beta > 1/2$, voir proposition 3.5. Pour $0 < \beta \leq 1/2$, ils obtiennent également une convergence en loi mais avec une autre normalisation, voir propositions 3.5 et 3.9.

On montre également que Z_β et son moment d'ordre 1 sont presque sûrement finis si $\beta > 1/2$. Dans ce cas, on obtient l'expression du moment d'ordre 1, retrouvant ainsi l'expression donnée par Fill et Kapur [89]. Si $0 < \beta \leq 1/2$ alors Z_β est presque sûrement infinie. On met ainsi en évidence une transition de phase en $\beta = 1/2$.

La proposition C permet en particulier de retrouver, pour $\beta = 1$ et $\beta = 2$, les convergences classiques, voir par exemple Aldous [9] et Janson [113] :

$$|\mathrm{T}_n|^{-\frac{3}{2}} \sum_{v\in \mathrm{T}_n} |\mathrm{T}_{n,v}| \xrightarrow[n\to+\infty]{p.s.} 2\int_0^1 B_{\mathrm{ex}}(s)ds$$

et

$$|\mathrm{T}_n|^{-\frac{5}{2}} \sum_{v\in \mathrm{T}_n} |\mathrm{T}_{n,v}|^2 \xrightarrow[n\to+\infty]{p.s.} 4\int_{0\leq s\leq t\leq 1} m_{B_{\mathrm{ex}}}(s,t)dsdt.$$

Nous terminons par les fluctuations associées à ce principe d'invariance :

Proposition D. *Pour tout $\beta \geq 1$, on a la convergence en loi suivante :*

$$\left(|\mathrm{T}_n|^{1/4}(Z_\beta^{(n)} - 2Z_\beta), Z_\beta^{(n)}\right) \xrightarrow[n\to+\infty]{(d)} \left(\sqrt{2Z_{2\beta}}G, 2Z_\beta\right),$$

où G est une variable aléatoire de loi normale centrée, réduite indépendante de l'excursion brownienne e.

1.5.2 Résultats pour les arbres simplement générés

Que se passe-t-il maintenant si l'arbre binaire à n nœuds internes sous le modèle de Catalan T_n est remplacé par un arbre simplement généré $\tau^{(p)}$ à p nœuds ?

Contexte

Soit $\tau^{(p)}$ un arbre simplement généré à p nœuds de fonction poids $\mathfrak{p}$. On suppose que $\mathfrak{p}$ est une probabilité critique sur $\mathbb{N}$, avec $0 < \mathfrak{p}(0) \leq \mathfrak{p}(1) + \mathfrak{p}(0) < 1$ et qui appartient au domaine d'attraction d'une loi stable symétrique d'exposant de Laplace $\psi(\lambda) = \kappa\lambda^\gamma$ avec $\gamma \in (1,2]$ et $\kappa > 0$, et de suite normalisée $(a_p, p \in \mathbb{N}^*)$. Pour $v \in \tau^{(p)}$, on note $\tau_v^{(p)}$ le sous-arbre de $\tau^{(p)}$ au dessus du nœud v.

Convergence d'une mesure aléatoire

De manière similaire à la mesure aléatoire pondérée A_n, on définit la mesure aléatoire finie non normalisée, pour tout $f \in \mathcal{B}([0,1])$ par :

$$\mathcal{A}_{\tau^{(p)}}(f) = \sum_{v\in\tau^{(p)}} |\tau_v^{(p)}| f\left(\frac{|\tau_v^{(p)}|}{|\tau^{(p)}|}\right) = \sum_{v\in\tau^{(p)}} |\tau_v^{(p)}| f\left(\frac{|\tau_v^{(p)}|}{p}\right).$$

On rappelle que H est le processus hauteur associé à l'arbre de Lévy continu, voir paragraphe 1.2.3. On obtient la convergence en loi de la mesure $\mathcal{A}_{\tau^{(p)}}$ correctement normalisée.

Théorème E. *Soit $\mathfrak{p}$ une loi de probabilité critique sur $\mathbb{N}$, avec $0 < \mathfrak{p}(0) \leq \mathfrak{p}(1) + \mathfrak{p}(0) < 1$, qui appartient au domaine d'attraction d'une loi stable symétrique d'exposant de Laplace $\psi(\lambda) = \kappa\lambda^\gamma$ avec $\gamma \in (1, 2]$ et $\kappa > 0$, et de suite normalisée $(a_p, p \in \mathbb{N}^*)$. Soit τ un arbre de Galton-Watson de loi de distribution $\mathfrak{p}$, et $\tau^{(p)}$ de même loi que τ conditionnellement à $\{|\tau| = p\}$. On a la convergence en loi suivante :*

$$\frac{a_p}{p^2}\mathcal{A}_{\tau^{(p)}} \xrightarrow[p\to+\infty]{(d)} \Phi_H,$$

où on a muni l'espace des mesures positives de la topologie de la convergence étroite et où la convergence a lieu le long de toutes sous-suites infinies de p telles que $\mathbb{P}(|\tau| = p) > 0$.

En particulier, on a la convergence en loi suivante : pour tout $f \in \mathcal{C}([0, 1])$,

$$\frac{a_p}{p^2} \sum_{v\in\tau^{(p)}} |\tau_v^{(p)}|\ f\left(\frac{|\tau_v^{(p)}|}{p}\right) \xrightarrow[p\to+\infty]{(d)} \Phi_H(f).$$

La preuve repose sur la convergence globale des arbres de Galton-Watson renormalisés vers des arbres de Lévy en utilisant la convergence du processus de contour, voir la partie 1.2.3 concernant les travaux de Le Gall et Le Jan [136], Duquesne et Le Gall [70] et Duquesne [69].

Remarque. Si $\mathfrak{p}$ a une variance σ^2 finie alors on peut choisir $a_p = \sqrt{p}$ et H égale à $(2/\sigma)B_{\text{ex}}$.

Exemples d'applications :

Comme pour les arbres binaires sous le modèle de Catalan, on peut appliquer le théorème E à la fonction f définie par $f(x) = x^{\beta-1}$ pour $x \in [0, 1]$ et $\beta \geq 1$.

Sous les conditions du théorème E, il existe une suite $(a_p, p \in \mathbb{N}^*$ tel que $\mathbb{P}(|\tau| = p) > 0)$ telle que l'on a la convergence en loi suivante :

$$\frac{a_p}{p^{\beta+1}} \sum_{v\in\tau} |\tau_v^{(p)}|^\beta \xrightarrow[p\to+\infty]{(\text{d})} Z_\beta^H = \Phi_H(x^{\beta-1}),\ \text{pour tout } \beta \geq 1. \tag{1.7}$$

Maintenant, si $\mathfrak{p}$ a une variance σ^2 finie alors on peut choisir $a_p = \sqrt{p}$ et H égale à $(2/\sigma)B_{\text{ex}}$. En particulier, en utilisant le fait que $\Phi_{cB_{\text{ex}}} = c\Phi_{B_{\text{ex}}}$ où B_{ex} est l'excursion brownienne normalisée, on obtient les convergences pour les modèles suivants (voir tableau 1.1 pour les différents choix de σ^2) et pour tout $\beta \geq 1$:

- Arbres de Cayley à p nœuds :

$$\frac{1}{p^{\beta+\frac{1}{2}}} \sum_{v\in\mathbf{t}} |\tau_v^{(p)}|^\beta \xrightarrow[p\to+\infty]{(\text{d})} 2\,\Phi_{B_{\text{ex}}}(x^{\beta-1}).$$

- Arbres de Catalan à p nœuds :

$$\frac{1}{p^{\beta+\frac{1}{2}}} \sum_{v\in\mathbf{t}} |\tau_v^{(p)}|^\beta \xrightarrow[p\to+\infty]{(\text{d})} \sqrt{2}\,\Phi_{B_{\text{ex}}}(x^{\beta-1}).$$

On retrouve également la convergence des fonctionnelles associées aux fonctions péages de type puissance pour les arbres binaires sous le modèle de Catalan (la convergence est ici plus faible puisqu'elle est établie pour tout $\beta \geq 1$ tandis qu'elle l'était pour tout $\beta > 0$ précédemment).

1.5.3 Extensions possibles

Transition de phase pour les arbres simplement générés

Dans le cas des arbres binaires sous le modèle de Catalan, nous observons une transition de phase en $\beta = 1/2$ pour les fonctionnelles additives liées à des fonctions péages monômes $b_n = n^\beta$ avec $\beta > 0$. Quand $\beta > 1/2$, la fonctionnelle normalisée converge presque sûrement vers une limite Z_β finie tandis que quand $0 < \beta \leq 1/2$, cette limite est infinie. Ainsi pour $0 < \beta \leq 1/2$ la convergence n'est plus pertinente, voir Fill et Kapur [89], partie 3.4 pour la normalisation adaptée.
On aimerait montrer une transition de phase similaire pour les fonctionnelles additives pour des fonctions péages monômes $b_n = n^\beta$ avec $\beta > 0$ liées à des arbres simplement générés On conjecture que sous les conditions du théorème E, la transition de phase est en $\beta = 1/\gamma$. En effet, on a montré la convergence en loi de ces fonctionnelles additives pour tout $\beta \geq 1$, vers une limite notée Z_β^H, voir (1.7). De plus, on montre que presque sûrement pour tout $0 < \beta \leq 1/\gamma$, Z_β^H est infinie tandis que presque sûrement, pour tout $\beta > 1/\gamma$, Z_β^H est finie, voir le lemme 3.17 du chapitre 3. Ainsi, pour prouver la transition de phase, il faudrait montrer que la convergence des fonctionnelles additives a également lieu pour $1/\gamma < \beta < 1$.

Fonctionnelles de coût dites asymétriques

Soit T_n un arbre binaire ayant n nœud internes. On aimerait s'intéresser à des fonctionnelles dites asymétriques dont la fonction péage dépendrait non plus seulement du cardinal de l'arbre entier mais du cardinal de ses sous-arbres de gauche et de droite.
Un exemple de telles fonctionnelles est donné par l'indice de déséquilibre (« imbalance parameter » en anglais) qui compte le nombre de nœuds qui ont leurs sous-arbres de gauche et de droite de même taille, voir Devroye [60] :

$$\sum_{v \in \mathrm{T}_n} \mathbf{1}_{\{|L_{n,v}| = |R_{n,v}|\}},$$

où $L_{n,v}$ est le sous-arbre de gauche de T_n au dessus du nœud v et $R_{n,v}$ est le sous-arbre de droite de T_n au dessus du nœud v.

Il serait donc naturel de s'intéresser à des fonctionnelles qui mesurent l'asymétrie des arbres. Pour $n \in \mathbb{N}^*$, on peut ainsi définir la mesure aléatoire pondérée B_n sur $[0,1]$ définie pour tout $f \in \mathcal{B}([0,1]^2)$ par :

$$B_n(f) = |\mathrm{T}_n|^{-2} \sum_{v \in \mathrm{T}_n} |L_{n,v}|\,|R_{n,v}|\, f\left(\frac{|L_{n,v}|}{|\mathrm{T}_n|}, \frac{|R_{n,v}|}{|\mathrm{T}_n|}\right),$$

On conjecture la convergence de cette mesure aléatoire pour les arbres binaires sous le modèle de Catalan.

CHAPITRE 2

FONCTIONNELLES DE GRAPHES ALÉATOIRES ÉCHANTILLONNÉS À PARTIR D'UN GRAPHON

2.1 Graphes

Notations préliminaires Dans tout ce chapitre, F et G désignent des graphes finis simples. Pour $n \in \mathbb{N}^*$, on note $[n] \coloneqq \{1, \dots, n\}$. On note $|B|$ le cardinal de l'ensemble B. On désigne par $\mathcal{N}(m, \sigma^2)$ une loi normale de moyenne m et de variance σ^2.

2.1.1 Notions sur les graphes

Un graphe simple (non orienté) G est un couple $(V(G), E(G))$ formé d'un ensemble $V(G)$ de $v(G)$ sommets et d'un sous-ensemble $E(G)$ d'arêtes choisies dans la collection des $\binom{v(G)}{2}$ paires non ordonnées de sommets. On note $e(G)$ le nombre d'arêtes du graphe G. Rappelons qu'un graphe simple est un graphe sans boucles (un sommet n'est jamais relié à lui même) et sans arêtes multiples (deux sommets sont reliés par au plus une arête). Un graphe est fini si son nombre de sommets (et donc d'arêtes) est fini. Dans ce cas, en numérotant les sommets de 1 à $v(G)$, on pourra identifier $V(G)$ à $[v(G)]$. On considérera dans la suite des graphes finis simples (et non orientés). On note $\mathcal{F}$ l'ensemble des graphes finis simples (et non orientés).

Un graphe fini simple G peut être caractérisé par son ensemble de sommets $V(G) = [v(G)]$ et sa matrice d'adjacence A de dimension $v(G) \times v(G)$ définie par :

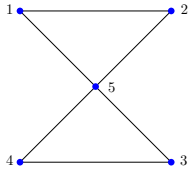
$$A(i,j) = \begin{cases} 1 & \text{si } \{i,j\} \in E(G) \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, A est une matrice symétrique binaire avec des zéros sur la diagonale. On peut également associer à une matrice d'adjacence d'un graphe G son image pixelisée (« pixel picture » en anglais) : il s'agit du carré unité que l'on subdivise en petits carrés de longueur $1/v(G)$; les 0 sont remplacés par des carrés blancs et les 1 sont remplacés par des carrés noirs.

Nous verrons que l'image pixelisée peut donner une intuition de la convergence d'une suite de grands graphes aléatoires, voir partie 2.1.3.

Le nombre maximal d'arêtes dans un graphe simple G à $v(G)$ sommets est $\binom{v(G)}{2}$. La densité ρ d'arêtes du graphe G est la fraction de ces arêtes qui sont réellement présentes :

$$\rho = \frac{e(G)}{\binom{v(G)}{2}} = \frac{2e(G)}{v(G)(v(G)-1)}.$$



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

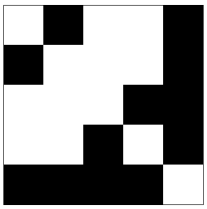


Figure 2.1 – Un graphe à 5 sommets, sa matrice d'adjacence et son image pixelisée.

On dira qu'une suite de graphes simples finis $(G_n : n \in \mathbb{N}^*)$ est dense (resp. creuse) si la suite de densités de ces graphes tend vers une constante strictement positive (resp. vers 0) quand n tend vers l'infini.

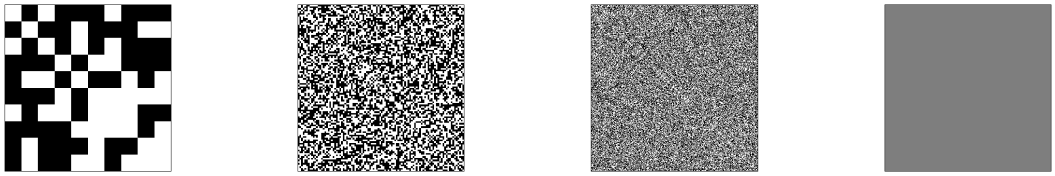
2.1.2 Graphes aléatoires

Nous allons maintenant considérer des graphes aléatoires. Les modèles de graphes aléatoires les plus simples sont ceux d'Erdős et Rényi, introduits en 1959 − 1960. Rappelons qu'un graphe d'Erdős-Rényi $G_n(\mathfrak{p})$ à n sommets et de paramètre $\mathfrak{p}$ avec $0 < \mathfrak{p} < 1$, est un graphe aléatoire ayant pour ensemble des sommets $[n] = \{1, \ldots, n\}$ et dont les arêtes existent de manière indépendante avec probabilité $\mathfrak{p}$. A partir des années 2000, d'autres modèles de graphes, plus complexes, ont été développés afin d'obtenir des graphes ayant des propriétés plus proches des réseaux réels. Parmi ces modèles figurent, par exemple, le modèle de configuration (modèle de graphes aléatoires où les degrés des sommets sont fixés) ou encore le modèle d'attachement préférentiel (modèle de graphes aléatoires dynamique, construits de manière récursive), voir Durrett [72] et van der Hofstad [187] pour de récents ouvrages de références.

2.1.3 Heuristique de la convergence d'une suite de graphes aléatoires

Dans cette partie, nous allons voir de manière informelle, quelle peut être la limite d'une suite de graphes au travers de deux exemples. Les exemples que nous développerons sont tirés de Borgs, Chayes, Lovász, Sós et Vesztergombi [37]. On considère une suite de graphes finis simples denses $(G_n : n \in \mathbb{N}^*)$. Quelle peut-être la limite de cette suite ?

Commençons par le modèle d'Erdős-Rényi. Soit $(G_n(1/2) : n \in \mathbb{N}^*)$ une suite de graphes où $G_n(1/2)$ est un graphe d'Erdős-Rényi de paramètre $1/2$ à n sommets. Si l'on regarde la suite d'images pixelisées associées aux matrices d'adjacences, on peut remarquer une convergence « graphique » de cette suite d'images vers le carré unité uniformément gris, comme l'illustre l'image 2.2 ci-dessous. L'origine est placée dans le coin en haut à gauche afin d'être en accord avec la convention de numérotation des éléments matriciels.

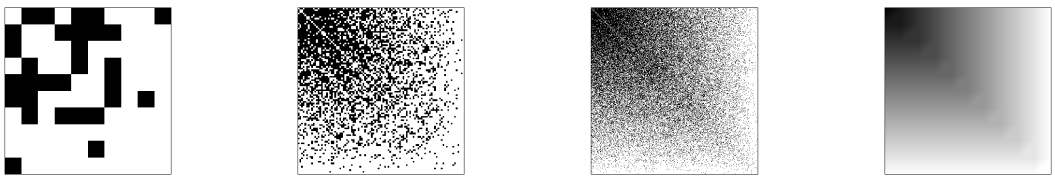


(a) $n = 10$ (b) $n = 100$ (c) $n = 1000$ (d) $W = \frac{1}{2}$

Figure 2.2 – Graphes d'Erdős-Rényi de paramètre $\mathfrak{p} = \frac{1}{2}$ et leur limite.

Ainsi, on peut conjecturer que la suite de graphes $(G_n(1/2) : n \in \mathbb{N}^*)$ « converge », en un sens que nous définirons dans la suite, vers le graphe infini dont la matrice d'adjacence correspond à la fonction constante égale à $1/2$ sur $[0,1]^2$.

Intéressons nous à un deuxième exemple. Considérons le modèle de graphes aléatoires d'attachement uniforme croissant que l'on notera $(GUA_n : n \in \mathbb{N}^*)$. Le graphe GUA_n ayant pour ensemble de sommets $[n]$ se construit de manière itérative. On commence par un seul sommet. À chaque étape un nouveau sommet est crée et chaque paire de sommets non encore connectés à l'étape précédente est reliée avec probabilité $1/k$ où k est le nombre de sommets à l'étape actuelle (voir figure 2.3). On remarque encore une fois une convergence graphique des images pixelisées quand le nombre de sommets augmente, vers la fonction de deux variables W définie par $W(x,y) = 1 - \max(x,y)$ pour tout $x, y \in [0,1]$.



(a) $n = 10$ (b) $n = 100$ (c) $n = 1000$ (d) $W(x,y) = 1 - \max(x,y)$

Figure 2.3 – Graphes aléatoires d'attachement uniforme croissant et leur limite

2.2 Des graphes aux graphons

2.2.1 Densités d'homomorphismes

Nous allons définir la notion d'homomorphismes qui permettra de donner une caractérisation de la convergence d'une suite de graphes denses. De nombreux paramètres de graphes peuvent s'exprimer à partir du nombre d'homomorphismes, voir Freedman, Lovasz et Schrijver [95] pour des exemples d'applications.

Soient $F, G \in \mathcal{F}$ deux graphes simples tels que $p = v(F)$ et $n = v(G)$. Dans la suite, il faut voir F comme un petit graphe et G comme un grand graphe. Un homomorphisme φ de F dans G est une application de $V(F) = [p]$ dans $V(G) = [n]$ qui préserve l'adjacence des arêtes. Autrement dit, c'est une application de $V(F)$ dans $V(G)$ telle que si $\{i,j\} \in E(F)$ alors $\{\varphi(i), \varphi(j)\} \in E(G)$. On note $\mathrm{Hom}(F,G)$ l'ensemble des homomorphismes de F dans G. Par exemple, si F est un triangle alors |Hom(F,G)| est le nombre de triangles dans G multiplié par 6 (permutations des sommets). On peut ainsi définir la densité d'homomorphismes de F dans G comme la quantité normalisée :

$$t(F,G) = \frac{|\mathrm{Hom(F,G)}|}{n^p},$$

qui est la probabilité qu'une application de $V(F)$ dans $V(G)$ choisie au hasard soit un homomorphisme.

De manière similaire, on note $\mathrm{Inj}(F,G)$ l'ensemble de homomorphismes injectifs et on lui associe sa densité :

$$t_{\mathrm{inj}}(F,G) = \frac{|\mathrm{Inj}(F,G)|}{A_n^p},$$

où pour tout $n \geq k \geq 1$, $A_n^k = n!/(n-k)!$. Les notations utilisées suivent celles de Borgs, Chayes, Lovász, Sós et Vesztergombi [40] (dans Bollobás et Riordan [32], t_{inj} est notée t).

Enfin, on définit l'ensemble des homomorphismes induits de F dans G, noté $\text{Ind}(F,G)$. Rappelons qu'un homomorphisme induit de F dans G est un homomorphisme injectif qui préserve également les relations de non-adjacence entre les arêtes. Autrement dit, c'est une application injective φ de $V(F)$ dans $V(G)$ telle que $\{i,j\} \in E(F)$ si et seulement si $\{\varphi(i),\varphi(j)\} \in E(G)$. On définit la densité d'homomorphismes induits par :

$$t_{\text{ind}}(F,G) = \frac{|\text{Ind}(F,G)|}{A_n^p}.$$

Il est facile de voir que les différentes notions d'homomorphismes sont reliées, voir par exemple Lovász et Szegedy [139], paragraphe 2.4. Pour $F, G \in \mathcal{F}$ deux graphes finis simples, on a :

$$|t_{\text{inj}}(F,G) - t(F,G)| \leq \frac{1}{n}\binom{p}{2}, \tag{2.1}$$

ainsi que

$$t_{\text{inj}}(F,G) = \sum_{F' \geq F} t_{\text{ind}}(F',G) \quad \text{et} \quad t_{\text{ind}}(F,G) = \sum_{F' \geq F} (-1)^{e(F')-e(F)} t_{\text{inj}}(F',G), \tag{2.2}$$

où $F' \geq F$ signifie que $V(F) = V(F')$ et $E(F) \subset E(F')$, i.e. F' parcourt l'ensemble des graphes simples obtenus à partir de F en ajoutant des arêtes.

Comptes de sous-graphes De nombreux auteurs étudient les comptes de sous-graphes plutôt que les densités d'homomorphismes. Deux graphes F_1 et F_2 sont isomorphes (ou F_2 est une copie de F_1) s'il existe un homomorphisme induit bijectif φ de $V(F_1)$ dans $V(F_2)$. L'application φ est aussi appelée un isomorphisme. Pour deux graphes F et G, on définit le nombre de copies de F dans G, i.e. le nombre de sous-graphes de G qui sont isomorphiques à F :

$$\chi_F(G) = \left|\left\{F' \subset G : F' \text{ est une copie de } F\right\}\right|, \tag{2.3}$$

où $F' \subset G$ signifie que $V(F) \subset V(G)$ et $E(F) \subset E(G)$. Un automorphisme d'un graphe F est un isomorphisme de graphe de F dans lui-même. On note $|\text{Aut}(F)| = \chi_F(F)$ son cardinal. On note K_n le graphe complet à n sommets. On a les relations suivantes entre les comptes de sous-graphes, les nombres d'homomorphismes injectifs et les densités d'homomorphismes, voir Bollobás et Riordan [32], Section 2.1 : soient $F, G \in \mathcal{F}$ deux graphes simples ayant p et n sommets respectivement et tels que $n \geq p \geq 1$. On a les égalités suivantes :

$$\chi_F(K_n) = \frac{A_n^p}{|\text{Aut}(F)|} \quad \text{et} \quad \chi_F(G) = \frac{A_n^p}{|\text{Aut}(F)|}\, t_{\text{inj}}(F,G).$$

2.2.2 Graphons

Dans cette partie, nous allons introduire la notation de graphons défini par Lovász et Szegedy [139] en 2006 qui apparaîtra dans la suite, comme l'objet limite adapté pour décrire les limites de suites de graphes denses. Nous étendrons également les définitions d'homomorphismes de graphes aux graphons.

On appelle graphon toute fonction W mesurable symétrique de $[0,1]^2$ dans $[0,1]$. On note $\mathcal{W}$ l'espace des graphons.

Construction d'un graphon à partir d'un graphe : Pour tout graphe simple G à n nœuds, il existe une manière naturelle de construire un graphon W_G associé au graphe G, voir Borgs, Chayes, Lovász, Sós et Vesztergombi [40], paragraphe 3.1. Pour ce faire, on divise l'intervalle $[0,1]$ en n intervalles $(J_i : 1 \leq i \leq n)$ de longueur $1/n$ où pour tout $i \in \{1,\dots,n\}$, $J_i = \left[\frac{i-1}{n}, \frac{i}{n}\right)$ et pour tout $x \in J_i$, $y \in J_j$, on définit W_G par :

$$W_G(x,y) = \begin{cases} 1 \text{ si } \{i,j\} \in E(G) \\ 0 \text{ sinon.} \end{cases}$$

Ainsi, le graphe de W_G est l'image pixelisée du graphe G. La fonction W_G est aussi appelée graphon empirique (« empirical graphon » en anglais).

Densités d'homomorphismes associées aux graphons : La correspondance entre les graphes simples et les graphons suggère qu'il est possible d'étendre de manière naturelle les densités d'homomorphismes de graphes aux graphons, voir Borgs, Chayes et Lovász [36]. Pour tout graphe simple F et pour tout graphon $W \in \mathcal{W}$, on définit

$$t(F,W) = t_{\text{inj}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i,x_j) \prod_{k\in V(F)} dx_k \tag{2.4}$$

et

$$t_{\text{ind}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i,x_j) \prod_{\{i,j\}\notin E(F)} (1-W(x_i,x_j)) \prod_{k\in V(F)} dx_k. \tag{2.5}$$

Ces formules peuvent être interprétées comme le « nombre » normalisé d'homomorphismes injectifs (resp. homomorphismes induits) de F dans un graphe ayant pour ensemble de sommets $[0,1]$ et de poids entre les arêtes donnés par W.

Pour tout graphe simple $F \in \mathcal{F}$ et tout graphe simple G, on a :

$$t(F,G) = t(F,W_G)\cdot$$

Les densités d'homomorphismes injectifs t_{inj} coïncident dans le cadre des graphons avec les densités d'homomorphismes t puisque l'on a $t_{\text{inj}}(F,W) = t(F,W)$, pour tout $W \in \mathcal{W}$ et $F \in \mathcal{F}$. En effet, intuitivement, les tirages de sommets avec ou sans remise coïncident quand le nombre de sommets est infini.

Pour terminer cette partie, nous introduisons la fonction de degré normalisée D d'un graphon W qui interviendra dans nos résultats du chapitre 4. Elle est définie pour tout $x \in [0,1]$ par :

$$D(x) = \int_0^1 W(x,y)dy.$$

En notant K_2 le sous-graphe complet à deux sommets (autrement dit, une arête), on remarque que l'on a :

$$t_{\text{inj}}(K_2,W) = \int_0^1 D(x)\,dx.$$

2.2.3 Convergence d'une suite de graphes denses

Dans cette section, nous avons choisi d'introduire deux notions de convergence de suite de graphes denses : la convergence associée aux densités d'homomorphismes (celle qui sera l'objet de notre étude) et la convergence métrique pour la distance de coupe. D'autres convergences équivalentes aux deux précédentes ont également été établies, voir Borgs, Chayes et Gamarnik [34] et Borgs, Chayes, Lovász, Sós et Vesztergombi [41].

Notons également que la théorie de limites de graphes est aussi intimement reliée à la théorie des tableaux partiellement échangeables de variables aléatoires étudiée par Aldous [12] et Hoover [110]. Le lecteur pourra se référer à Diaconis et Janson [61] et Lovász [138], section 11.3.3 pour plus de détails.

Densités d'homomorphismes et convergence

Dans ce paragraphe, nous allons définir la première notion de convergence de suites de graphes denses introduite par Lovász et Szegedy [139] en 2006. Elle fait intervenir les densités d'homomorphismes définies dans la section 2.2.1. Cette convergence est également appelée convergence à gauche, voir Lovász [138], chapitre 11.

Lovász et Szegedy [139] ont énoncé la définition suivante :

Définition. *Une suite de graphes finis simples* $(G_n : n \in \mathbb{N}^*)$ *est dite convergente si la suite* $(t_{\text{inj}}(F, G_n) : n \in \mathbb{N}^*)$ *admet une limite pour tout* $F \in \mathcal{F}$.

Par les formules qui relient t, t_{inj} et t_{ind}, c'est équivalent à demander que la suite $(t(F, G_n) : n \in \mathbb{N}^*)$ converge pour tout $F \in \mathcal{F}$ ou que la suite $(t_{\text{ind}}(F, G_n) : n \in \mathbb{N}^*)$ converge pour tout $F \in \mathcal{F}$.

Le résultat principal de Lovász et Szegedy [139] est le suivant :

Theorem (Lovász et Szegedy (2006))**.** *Pour toute suite de graphes convergente* $(G_n : n \in \mathbb{N}^*)$, *il existe un graphon* $W \in \mathcal{W}$ *tel que*

$$\lim_{n \to +\infty} t(F, G_n) = t(F, W), \quad \forall F \in \mathcal{F}. \tag{2.6}$$

La preuve originale de Lovász et Szegedy de ce théorème utilise le lemme de régularité de Szemerédi [182] et le théorème de convergence des martingale. Deux autres preuves ont également été établies par Elek et Szegedy [73] en utilisant la théorie des ultra produits et ultra limites et par Diaconis et Janson [61] grâce à la théorie des variables échangeables.

Borgs, Chayes et Lovász [36] on montré l'unicité de la limite W à une bijection préservant la mesure près. Autrement dit, pour toute autre fonction limite $W' \in \mathcal{W}$, il existe une application bijective $\varphi : [0,1] \to [0,1]$ telle que φ et son inverse préservent la mesure et telle que $W(x,y) = W'(\varphi(x), \varphi(y))$ pour presque tout $x, y \in [0,1]$.

On dira qu'une suite de graphes $(G_n : n \in \mathbb{N}^*)$ converge vers un graphon $W \in \mathcal{W}$ si (2.6) a lieu.

Lovász et Szegedy ont également établi la réciproque du théorème précédent :

Proposition (Lovász et Szegedy (2006))**.** *Pour tout graphon* $W \in \mathcal{W}$, *il existe une suite de graphes* $(G_n : n \in \mathbb{N}^*)$ *telle que*

$$\lim_{n \to +\infty} t(F, G_n) = t(F, W), \quad \forall F \in \mathcal{F}.$$

Nous verrons dans la sous-section 2.3 un exemple de construction d'une telle suite de graphes $(G_n : n \in \mathbb{N}^*)$.

Distance de coupe et convergence

La deuxième notion de convergence proposée par Borgs, Chayes, Lovász, Sós et Vesztergombi [38, 40] est une convergence métrique définie à partir de la distance de coupe. La

distance de coupe est basée sur une norme de coupe notée $\|.\|_\square$ introduite par Frieze et Kannan [96] en 1999. Pour A une matrice réelle carrée de taille $n \times n$ avec $n \in \mathbb{N}^*$, on définit sa norme de coupe par :

$$\|A\|_\square = \frac{1}{n^2} \max_{S,T \subseteq [n]} \left| \sum_{i \in S, j \in T} A(i,j) \right|.$$

Il est naturel d'étendre cette norme aux fonctions mesurables symétriques et bornées de $[0,1]^2$ dans $\mathbb{R}$. On définit la norme de coupe notée $\|.\|_\square$, pour toute fonction mesurable symétrique et bornée U de $[0,1]^2$ dans $[0,1]$ par :

$$\|U\|_\square = \sup_{S,T \subseteq [0,1]} \left| \int_{S \times T} U(x,y) dx\, dy \right|,$$

où le supremum est pris sur l'ensemble des sous-ensembles mesurables S et T de $[0,1]$.

La distance de coupe sur l'espace des graphons, notée $\delta_\square$, est définie pour tous graphons $W_1, W_2 \in \mathcal{W}$ par :

$$\delta_\square(W_1, W_2) = \inf_{\phi:[0,1]\to[0,1]} \|W_1 - W_2^\phi\|_\square,$$

où l'infimum est pris sur toutes les fonctions bijectives $\phi : [0,1] \to [0,1]$ telle que ϕ et son inverse préservent la mesure et où W_2^ϕ est défini par $W_2^\phi(x,y) = W(\phi(x), \phi(y))$, pour tout $x, y \in [0,1]$. La distance $\delta_\square$ est en fait une pseudo-distance. En effet, deux graphons différents peuvent être à distance nulle pour la distance de coupe. En identifiant tous les noyaux dont la distance de coupe est nulle, on obtient l'espace quotient noté $\tilde{\mathcal{W}}$ qui est l'ensemble des graphons non étiquetés.

Nous énonçons maintenant le théorème établi par Borgs, Chayes, Lovász, Sós et Vesztergombi [38] en 2006 qui donne l'équivalence entre la convergence métrique et la convergence au sens des densités d'homomorphismes, justifiant ainsi l'introduction de la distance de coupe : une suite de graphes simples finis $(G_n : n \in \mathbb{N}^*)$ est convergente si et seulement si elle est de Cauchy pour la distance de coupe $\delta_\square$. De plus, la suite de graphes simples finis $(G_n : n \in \mathbb{N}^*)$ converge vers $W \in \mathcal{W}$ si et seulement si $\lim_{n\to+\infty} \delta_\square(W_{G_n}, W) = 0$.

Nous terminons ce paragraphe sur une propriété de l'espace $\tilde{\mathcal{W}}$ muni de la distance de coupe. L'espace $(\tilde{\mathcal{W}}, \delta_\square)$ est compact, voir Lovász et Szegedy [139] ou Lovász [138], partie 9.3. En particulier, $(\tilde{\mathcal{W}}, \delta_\square)$ est un espace métrique complet. En combinant ce fait et le résultat précédent sur l'équivalence des convergences, on en déduit que l'espace des graphons est l'espace complété des graphes finis avec la distance de coupe.

2.3 Échantillonnage de graphes aléatoires à partir d'un graphon

Soit $W \in \mathcal{W}$. On définit un modèle de graphes aléatoires générés à partir du graphon W. On génère un graphe aléatoire $G_n(W)$ à n sommets avec $n \in \mathbb{N}^*$ à partir de W de la manière suivante : on échantillonne une suite de variables aléatoires $X = (X_i : i \in \mathbb{N}^*)$ indépendantes de loi uniforme sur $[0,1]$ puis, étant donnée cette suite, on met une arête entre les sommets i et j avec $i, j \in [n]$ avec probabilité $W(X_i, X_j)$. Pour une suite donnée X, on répète l'opération de manière indépendante pour chacune des paires $(i,j) \in [n]^2$ avec $i < j$. Le processus ainsi décrit repose donc sur deux niveaux d'aléa.

Remarquons que les graphes aléatoires $G_n(W)$ sont une généralisation des graphes aléatoires d'Erdős-Rényi $G_n(\mathfrak{p})$ obtenus en prenant le graphon constant égale à $\mathfrak{p}$ avec $0 < \mathfrak{p} < 1$.

Il a été prouvé par Lovász et Szegedy [139] en 2006, que la suite de graphes aléatoires $(G_n(W) : n \in \mathbb{N}^*)$ converge presque sûrement vers le graphon choisi W. Autrement dit :

Proposition (Lovász et Szegedy (2006)). *Pour tout graphe simple $F \in \mathcal{F}$, on a presque sûrement :*

$$\lim_{n\to+\infty} t_{\text{inj}}(F, G_n(W)) = t(F, W). \tag{2.7}$$

La preuve repose sur l'égalité $\mathbb{E}\left[t_{\text{inj}}(F, G_n(W))\right] = t(F, W)$, un lemme de concentration sur les graphons (conséquence directe de l'inégalité d'Azuma) et le lemme de Borel-Cantelli, voir Lovász [138], partie 11.4.1.

Dans le cas d'Erdős-Rényi, i.e. quand le graphon W est constant, on retrouve la conjecture graphique énoncée dans la sous-section 2.1.3 : la suite de graphes d'Erdős-Rényi de paramètre $1/2$ (resp. $\mathfrak{p}$ avec $0 < \mathfrak{p} < 1$) converge vers le graphon constant $W \equiv 1/2$ (resp. $W \equiv \mathfrak{p}$).

Nous allons maintenant nous intéresser aux fluctuations associées à ce principe d'invariance. Nous étudierons dans le chapitre 4, le cas où W est non constant, mais historiquement les fluctuations ont d'abord été montrées dans le cas où W est constant (autrement dit, pour le modèle d'Erdős-Rényi). C'est ce dernier cas que nous allons commencer par développer.

Cas particulier du graphon constant

Au tournant des années 80 et 90, Nowicki [159] et Janson et Nowicki [118] ont été les premiers à montrer que la vitesse de convergence associée à la convergence presque sûre est d'ordre n. Leurs preuves s'appuyaient sur la théorie des U-statistiques.

Proposition (Nowicki (1989), Janson et Nowicki (1991)). *Pour tout $F \in \mathcal{F}$ ayant e arêtes, on a la convergence en loi suivante :*

$$n\left(t_{inj}(F, G_n(\mathfrak{p})) - \mathfrak{p}^e\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, 2e^2\mathfrak{p}^{2e-1}(1-\mathfrak{p})\right).$$

Janson et Nowicki ont aussi obtenu la convergence des densités d'homomorphismes induits ainsi que la normalité asymptotique de vecteurs de comptes de sous-graphes et de comptes de sous-graphes induits. Dans le cas particulier de la loi jointe des comptes des arêtes, triangles et des graphes étoiles à 2 feuilles (« two-stars » en anglais), Reinert et Röllin [167], proposition 2, ont obtenu des bornes sur l'approximation. Par des méthodes de calculs de Malliavin, Krokowski et Thäle [132] ont généralisé le résultat de Reinert et Röllin (dans un espace de probabilité différent) et ont obtenu la vitesse de convergence associée au théorème central limite multidimensionnel de Jason et Nowicki [118]. Féray, Méliot et Nikeghbali [83] se sont également intéressés aux densités d'homomorphismes en lien avec la convergence « mod-gaussian ».

Chatterjee et Varadhan [46] ont quant à eux, développé une théorie de grandes déviations pour les graphes d'Erdős-Rényi en lien avec les graphons, voir aussi Chatterjee [44].

L'étude asymptotique des comptes normalisés de sous-graphes a également été faite dans le cas où le paramètre $\mathfrak{p}$ des graphes d'Erdős-Rényi dépend de n. Grâce à la méthode des moments, Ruciński [174] a déterminé toutes les valeurs de $\mathfrak{p}$ telles que les comptes normalisés de sous-graphes convergent en loi vers une loi normale. Nowicki et Wierman [161] ont appliqué la théorie des U-statistiques pour obtenir des résultats similaires mais dans un cadre plus faible. En utilisant la méthode de Stein, Barbour, Karoński et Ruciński [18] ont complété ces résultats en donnant des bornes sur l'erreur dans le théorème central limite. Dans le cas particulier du nombre de triangles, Gilmer et Kopparty [100] ont prouvé un théorème central limite local.

Cas où W n'est pas constant

Dans le cas où W est non constant, Féray, Méliot et Nikeghbali [84], théorème 21, ont établi fin 2017, alors que nous travaillions déjà sur ce sujet, que la vitesse de convergence associée au principe d'invariance est d'ordre $\sqrt{n}$.

Proposition (Féray, Méliot et Nikeghbali (2017)). *Pour tout $F \in \mathcal{F}$, on a la convergence en loi suivante :*

$$\sqrt{n}\left(t_{\mathrm{inj}}(F, G_n) - t(F, W)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma(F)^2\right),$$

où

$$\sigma(F)^2 = \sum_{q,q' \in V(F)} t\big((F \bowtie F)(q, q'), W\big) - v(F)^2\, t(F, W)^2$$

et $(F \bowtie F')(q, q')$ est l'union disjointe de deux graphes simples finis F and F' où l'on a identifié les sommets $q \in F$ et $q' \in F'$ (voir figure 2.4).

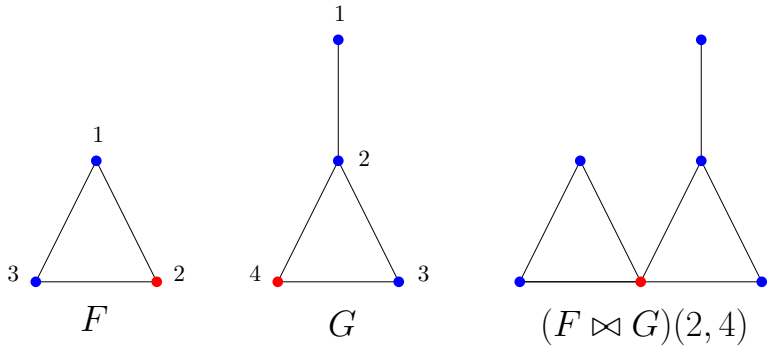


Figure 2.4 – Exemple de deux graphes connectés par deux sommets.

Remarque. Quand le graphon $W \equiv \mathfrak{p}$ est constant, il est facile de voir que la variance du théorème central limite énoncé par Féray, Méliot et Nikeghbali [84] est nulle, confirmant ainsi que la vitesse en $\sqrt{n}$ n'est plus pertinente dans ce cas.

Autres résultats de convergence

En utilisant la méthode de Stein, Fang et Röllin [81] ont obtenu la vitesse de convergence du théorème central limite multidimensionnel associé à la loi jointe des comptes normalisés d'arêtes et de cycles de longueur 4. Cela leur a notamment permis d'établir un intervalle de confiance pour tester si un graphe donné G provient du modèle d'Erdős-Rényi ou bien d'un modèle de W-graphes aléatoires avec un W non constant. Maugis, Priebe, Olhede et Wolfe [143] ont donné un théorème central limite pour les comptes de sous-graphes observés parmi une collection finie de W-graphes aléatoires générés à partir d'un même graphon W ou de plusieurs graphons et quand le nombre d'observations de l'échantillon augmente mais que le nombre de sommets de chaque graphe observé reste fini. Ceci leur permet de tester si les graphes observés proviennent d'un graphon particulier.

On peut également s'intéresser à des suites de graphons qui tendent vers 0. Cela mène à des approximations de Poisson des comptes de sous-graphes. Ainsi, Coulson, Gaunt et Reinert [51], corollaire 4.1, ont utilisé la méthode de Stein pour établir des approximations de Poisson pour la loi du nombre de sous-graphes qui sont isomorphiques à un graphe strictement équilibré donné.

2.4 Résultats du Chapitre 4

Dans cette partie, nous allons synthétiser les résultats du Chapitre 4 qui contient l'article [53] reporté sans modifications.

Contexte

Soit $W \in \mathcal{W}$ un graphon et $(G_n(W) : n \in \mathbb{N}^*)$ la suite de W-graphes aléatoires associée à W, voir partie 2.3.

Espaces fonctionnels

Soit $I = [0, 1]$. On note $\mathcal{B}(I)$ (resp. $\mathcal{B}^+(I)$) l'espace des fonctions mesurables de I dans $\mathbb{R}$ (resp. dans $\mathbb{R}^+$). On note $\mathcal{C}(I)$ l'espace des fonctions continues de I dans $\mathbb{R}$. Pour $f \in \mathcal{B}(I)$, on note $\| f \|_\infty$ la norme infinie de f sur I. On note $\mathcal{C}^k(I)$ l'ensemble des fonctions de I dans $\mathbb{R}$ qui ont leur k-ième dérivée continue sur I.

Nous allons dans un premier temps, nous intéresser au comportement asymptotique de la fonction de répartition empirique (FDR) des degrés associée à cette suite de graphes. Enfin, nous considérerons les asymptotiques des densités d'homomorphismes de la suite $(G_n(W) : n \in \mathbb{N}^*)$.

2.4.1 Fonction de répartition empirique des degrés

La littérature concernant les propriétés des graphes et en particulier des degrés de leurs sommets est très importante, voir par exemple Newman, Barabási et Watts [155], chapitre 3, Newman [158], chapitre 3 ou Bollobás [29]. Pour étudier la structure de grands graphes aléatoires, de nombreuses caractéristiques peuvent être analysées : nombre et taille des composantes, points isolés, degré d'un sommet, distance entre les sommets, étude des petits graphes (par exemple des triangles, arbres, cliques, ...) ou encore des notions autour de la centralité (« betweeness centrality » en anglais).

La suite des degrés (ou encore loi des degrés) est une des propriétés fondamentales des graphes, c'est celle qui nous intéressera dans la suite. On rappelle que le degré d'un sommet est égal au nombre d'arêtes qui lui sont rattachées.

Étude des degrés, graphes et graphons

Commençons par rappeler la notion de loi des degrés d'un graphe. Soit $G \in \mathcal{F}$. On définit p_k comme la fraction de sommets du graphe G qui sont de degrés k, pour tout $k \in [\![0, v(G)-1]\!]$. Les quantités p_k représentent la loi des degrés du graphe G. La valeur $(p_k : k \in [\![0, v(G) - 1]\!])$ peut également être vue comme la probabilité qu'un sommet choisi au hasard dans le graphe soit de degré k. Une autre notion qui contient essentiellement la même information que la loi des degrés est la suite des degrés d'un graphe. Insistons sur le fait que la loi des degrés (ou la suite des degrés) ne donne pas, en règle générale, la structure complète d'un graphe. On rappelle également que les degrés des réseaux réels (comme internet ou les réseaux sociaux) suivent une loi de puissance (« power-law » en anglais) : la plupart des sommets ont des petits degrés mais certains sommets ont des très grands degrés, voir Bollobás [29] ou Newman [154]. Ce n'est pas le cas des premiers modèles de graphes aléatoires, comme les graphes d'Erdős-Rényi dont la loi du degré de chaque sommet est une loi binomiale. Le modèle d'Erdős-Rényi apparaît donc inadapté pour modéliser les réseaux réels.

Pour cette raison entre autres, d'autres modèles de graphes, plus sophistiqués, ont été créés afin d'obtenir non plus des graphes où les degrés sont proches de lois binomiales mais de lois arbitraires données. En 1980, Bollobás [28] a été le premier à construire de tels graphes,

voir aussi Molloy et Reed [147], Newman, Strogatz and Watts [156, 157] ainsi que les ouvrages plus généraux de Bollobás [29], Durrett [72] (chapitre 3) et van der Hofstad [187]. Plusieurs algorithmes existent pour créer de tels graphes. Citons par exemple le modèle de configuration introduit en 1995 par Molloy et Reed [147] qui consiste à apparier des paires de sommets de manière uniforme, les algorithmes basés sur les listes d'adjacence et le théorème d'Havel-Hakami, voir Tinhofer [184], les algorithmes par méthode de Monte-Carlo par chaînes de Markov ou encore les algorithmes séquentiels, voir Blitzstein et Diaconis [26].

Afin d'obtenir des modèles encore plus proches des réseaux réels, on peut également construire des modèles statistiques plus généraux de graphes vérifiant non plus seulement une propriété (comme le nombre de sommets ou la loi des degrés) mais un ensemble de propriétés. C'est le cas du modèle de graphes exponentiels (modèle statistique qui donne une plus grande probabilité aux graphes qui correspondent le mieux aux caractéristiques observées) introduit par Holland et Leinhardt [107] en 1981, voir aussi Strauss [180], Frank et Strauss [94], Anderson, Wasserman et Crouch [14], Park et Newman [162] ou encore Robins, Pattison, Kalish et Lusher [170]. Il est également possible de construire des graphes par partitionnement des données (« clustering » en anglais), où l'ensemble des sommets des graphes est divisé en un certain nombre de blocs ou classes. Un des plus populaires est le modèle à blocs stochastiques (« stochastic bloc model » (SBM) en anglais) introduit en 1983 par Holland, Laskey et Leinhardt [106] et développé ensuite par Nowicki and Snijders [160]. Il décrit la probabilité d'apparition d'une arête entre deux nœuds uniquement en fonction des classes auxquelles ils appartiennent. Ce modèle est adapté pour analyser des petits graphes mais ne permet pas d'analyser les propriétés fines de très grands graphes. Des modèles non paramétriques de graphes aléatoires ont été développés comme le modèle de graphes aléatoires générés par un graphon. Ce dernier peut d'ailleurs être vu comme une généralisation du modèle à blocs stochastiques en prenant un graphon constant par morceaux. Ce lien est utilisé de manière déterminante pour l'estimation non-paramétrique des graphons.

L'estimation des graphons, consiste à retrouver, à partir de l'observation d'un ou plusieurs graphes, le modèle de graphon à partir duquel est engendré le ou les graphes. La littérature autour de la modélisation non-paramétrique des graphons est très riche, voir par exemple Airoldi, Costa et Chan [5], Wolfe et Olhede [194], Borgs, Chayes et Smith [39], Gao, Lu et Zhou [97], Latouche et Robin [133], Klopp, Tsybakov et Verzelen [124] ou Klopp et Verzelen [125].

Par ailleurs, des études paramétriques ont été développées en se basant sur la loi des degrés observée. Ainsi, Chatterjee, Diaconis et Sly [45] ont obtenu la limite de graphes aléatoires ayant une suite de degrés donnée vers un graphon dont la fonction degré est la limite de la suite de degrés. Ils ont utilisé le modèle de graphes exponentiels introduit par Holland et Leinhardt [107] en 1981. Bickel, Chen et Levina [25] ont montré, entre autres, que la FDR des degrés de graphes aléatoires échantillonnés à partir d'un graphon W converge en loi vers la fonction degré du graphon W.

Comportement asymptotique de la FDR empirique des degrés

Nous allons maintenant énoncer nos résultats concernant l'étude asymptotique de la FDR empirique des degrés d'une suite de graphes échantillonnés à partir d'un graphon. On rappelle que la fonction degré $D = (D(x) : x \in [0,1])$ du graphon W est définie par :

$$D(x) = \int_0^1 W(x,y)\,dy.$$

Remarque. Comme un graphon est défini à une bijection préservant la mesure près, il existe toujours une version équivalente au graphon pour lequel la fonction degré est croissante. Si la fonction degré est croissante alors cette version est unique dans L^1, voir Bickel et Chen [24]. C'est cette version dite « canonique » que nous considérerons dans la suite.

On considère la FDR empirique $\Pi_n = (\Pi_n(y) : y \in [0,1])$ de la suite des degrés normalisée du graphe $G_n(W)$ définie par

$$\Pi_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{D_i^{(n)} \leq D(y)\right\}},$$

où $nD_i^{(n)}$ est le degré du sommet i dans $G_n(W)$.

Remarque. Lorsque D est strictement croissante, D^{-1} est bien définie et pour tout $y \in [0,1]$, on a :

$$\Pi_n(D^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{D_i^{(n)} \leq y\right\}}.$$

On note Id l'application identité sur $[0,1]$. À l'image du théorème de Glivenko-Cantelli qui fournit la convergence presque sûre uniforme de la suite des fonctions de répartition empiriques d'un échantillon de variables aléatoires indépendantes et identiquement distribuées, nous obtenons presque sûrement, la convergence uniforme de la FDR empirique des degrés vers la fonction identité sur $[0,1]$.

Théorème A. *On suppose que D est strictement croissante sur $[0,1]$. Alors on a la convergence presque sûre suivante :*

$$\|\Pi_n - \mathrm{Id}\|_\infty \xrightarrow[n\to+\infty]{p.s.} 0.$$

Dans l'esprit du théorème de Kolmogorov-Smirnov qui fournit, sous une hypothèse supplémentaire, la vitesse de convergence associée au théorème de Glivenko-Cantelli, nous obtenons les fluctuations associées au théorème A en ajoutant des conditions sur W et D :

$$W \in \mathcal{C}^3([0,1]^2),\ D' > 0,\ W < 1 \text{ et } D > 0. \tag{2.8}$$

Théorème B. *On suppose que W satisfait la condition (2.8). Alors on a la convergence des lois finies dimensionnelles suivante :*

$$\left(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1)\right) \xrightarrow[n\to+\infty]{(lfd)} \chi,$$

où $\chi = (\chi_y : y \in (0,1))$ est un processus gaussien centré défini, pour tout $y \in (0,1)$ par :

$$\chi_y = \int_0^1 (\rho(y,u) - \bar{\rho}(y))dB_u,$$

avec $B = (B_u, u \geq 0)$ un mouvement brownien standard, et $(\rho(y,u) : u \in [0,1])$ et $\bar{\rho}(y)$ défini pour $y \in (0,1)$ par :

$$\rho(y,u) = \mathbf{1}_{[0,y]}(u) - \frac{W(y,u)}{D'(y)} \quad \text{et} \quad \bar{\rho}(y) = \int_0^1 \rho(y,u)du.$$

Remarque. On obtient également l'expression explicite du noyau de covariance du processus gaussien χ (voir remarque 4.24) ainsi que celle de sa variance donnée pour tout $y \in (0,1)$, par la formule suivante :

$$\Sigma(y,y) = y(1-y) + \frac{1}{D'(y)^2}\left(\int_0^1 W(y,x)^2dx - D(y)^2\right) + \frac{2}{D'(y)}\left(D(y)y - \int_0^y W(y,x)dx\right).$$

La preuve de ce théorème repose sur des développements uniformes d'Edgeworth pour des variables binomiales (voir Bhattacharya et Rao [22], Nagaev, Chebotarev et Zolotukhin [150] ou Uspensky [186]) et sur la méthode de Stein pour des vecteurs de variables binomiales de dimension 2 (voir Bentkus [20]).

Pour la convergence du processus $(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1))$ pour la topologie de Skorokhod, voir partie 2.4.3.

2.4.2 Densités d'homomorphismes de graphes partiellement étiquetés

Dans cette partie, nous allons étendre la notion de densité d'homomorphismes de graphes à des graphes partiellement étiquetés. Nous nous intéresserons ensuite aux résultats asymptotiques de mesures aléatoires construites à partir de ces densités d'homomorphismes pour des graphes aléatoires générés à partir d'un graphon.

Notations préliminaires Soit $n \in \mathbb{N}^*$ et $k \in [n]$. On définit l'ensemble $\mathcal{S}_{n,k}$ des $[n]$-mots de longueur k dont tous les caractères sont distincts, voir (4.7) pour une définition exacte. On a $|\mathcal{S}_{n,k}| = A_n^k = n!/(n-k)!$.

Homomorphismes de graphes partiellement étiquetés

La notion d'homomorphismes de graphes s'étend de manière naturelle aux graphes partiellement étiquetés. Un graphe partiellement étiqueté est un graphe simple fini dont certains de ses sommets sont étiquetés par différents entiers.

Soient $F, G \in \mathcal{F}$ deux graphes simples tels que $V(F) = [p]$ et $V(G) = [n]$. On suppose que $n \geq p > k \geq 1$. Soient $\ell \in \mathcal{S}_{p,k}$ l'ensemble des k sommets étiquetés de F et $\alpha \in \mathcal{S}_{n,k}$ l'ensemble des k sommets étiquetés de G. On définit l'ensemble $\mathrm{Inj}(F^\ell, G^\alpha)$ des homomorphismes injectifs f de F dans G tels que $\mathrm{f}(\ell_i) = \alpha_i$ pour tout $i \in [k]$, ainsi que sa densité :

$$t_{\mathrm{inj}}(F^\ell, G^\alpha) = \frac{|\mathrm{Inj}(F^\ell, G^\alpha)|}{A_{n-k}^{p-k}}.$$

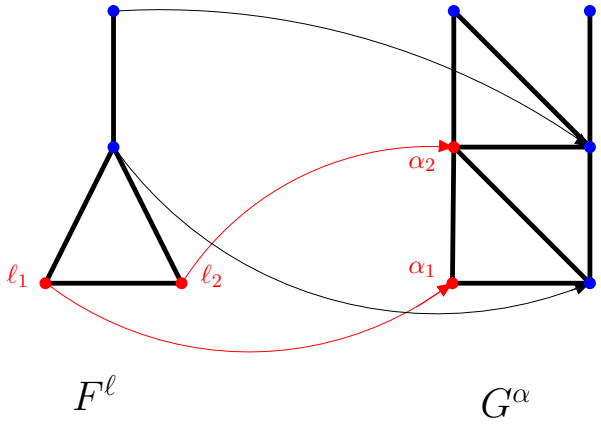


Figure 2.5 – Exemple d'homomorphisme injectif pour des graphes partiellement étiquetés..

Remarque. Il est facile de voir, en sommant sur tous les étiquetages de k sommets de G, que l'on a l'égalité suivante :

$$t_{\mathrm{inj}}(F, G) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\mathrm{inj}}(F^\ell, G^\alpha). \tag{2.9}$$

Une probabilité aléatoire

On définit la probabilité aléatoire $\Gamma_n^{F,\ell}$ sur $([0,1],\mathcal{B}([0,1]))$, où $\mathcal{B}([0,1])$ est la tribu borélienne sur $[0,1]$, par : pour tout $g \in \mathcal{B}^+([0,1])$:

$$\Gamma_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} g\left(t_{\text{inj}}(F^\ell, G_n^\alpha)\right). \tag{2.10}$$

Remarque. Afin de simplifier les résultats de convergence, la mesure a été définie pour des fonctions g d'une variable mais on peut également la définir pour des fonctions g à d variables avec $d \geq 1$ en considérant non plus un seul graphe simple F mais une suite de d graphes simples $F = (F_m : 1 \leq m \leq d) \in \mathcal{F}^d$. Les résultats multidimensionnels se trouvent dans le chapitre 4.

Résultats asymptotiques

Nous avons la convergence presque sûre suivante :

Théorème C. *La suite* $\left(\Gamma_n^{F,\ell}(dx) : n \in \mathbb{N}^*\right)$ *de mesures aléatoires sur* $[0,1]$ *converge presque sûrement pour la topologie de la convergence étroite de mesures sur* $[0,1]$ *vers la probabilité déterministe* $\Gamma^{F,\ell}(dx)$.

Afin de simplifier la lecture, nous avons choisi de ne pas donner l'expression exacte de la probabilité limite $\Gamma^{F,\ell}$ qui s'exprime en fonction des densités d'homomorphismes de graphes partiellement étiquetés dans des graphons (qui sont des généralisations naturelles des densités d'homomorphismes de graphes simples dans des graphons, voir la partie 4.2.3 du chapitre 4 pour une définition précise), voir l'équation (4.42) du chapitre 4.

En particulier, ce théorème permet d'obtenir les résultats suivants :

- Par le théorème de Portmanteau, on a presque sûrement, pour toute fonction $g \in \mathcal{C}([0,1])$, $\lim_{n\to+\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.
- Si on prend $g = \text{Id}$ dans (2.10), on retrouve la convergence donnée par (2.7) puisque l'on a d'après (2.9) :

  $$t_{\text{inj}}(F, G_n) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\text{inj}}(F^\ell, G_n^\alpha),$$

  et $\Gamma^{F,\ell}(\text{Id}) = t(F, W)$, voir remarque 4.8 du chapitre 4.
- Si on prend $g = \mathbf{1}_{[0,D(y)]}$ avec $y \in (0,1)$ et $F = K_2$ dans (2.10), on a, grâce à l'expression de $\Gamma^{F,\ell}$ donnée dans la remarque 4.8 du chapitre 4, avec $\bullet$ un des sommets de K_2, que :

  $$\Gamma_n^{K_2,\bullet}(g) = \Pi_n(y) \quad \text{et} \quad \Gamma^{K_2,\bullet}(g) = y.$$

  Si on suppose que D est strictement croissante sur $(0,1)$ alors le théorème C implique la convergence presque sûre de $\Pi_n(y)$ vers y pour tout $y \in (0,1)$. En utilisant le théorème de Dini, on obtient la convergence presque sûre de Π_n vers la fonction identité sur $[0,1]$ pour la norme uniforme, voir remarque 4.22 du chapitre 4.

Nous obtenons également les fluctuations associées à cette convergence presque sûre.

Théorème D. *Pour toute fonction* $g \in \mathcal{C}^2([0,1])$, *on a la convergence en loi suivante :*

$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$

avec $\sigma^{F,\ell}(g)^2 = \text{Var}\left(\mathcal{U}_g^{F,\ell}\right)$ *et* $\mathcal{U}_g^{F,\ell}$ *définie par l'équation (4.43) du chapitre 4.*

Pour $g = \mathrm{Id}$, on retrouve le théorème central limite énoncé par Féray, Méliot et Nikeghbali [84].

Par contre, comme $g = \mathbf{1}_{[0,D(y)]}$ n'appartient pas à l'ensemble $\mathcal{C}^2([0,1])$, on ne peut pas appliquer directement le théorème précédent (avec $F = K_2$ et $k = 1$) pour obtenir la convergence en loi du processus $(\sqrt{n}(\Pi_n(y) - y) : y \in [0,1])$ vers $(\chi(y) : y \in [0,1])$ donnée dans le théorème B. C'est pour cette raison que d'autres techniques ont dû être mises en place pour démontrer le théorème B.

Comme corollaire immédiat du théorème D énoncé en dimension supérieure, voir le théorème 4.11 du chapitre 4, on établit la convergence en lois finies-dimensionnelles (en fait, du processus puisque l'ensemble $\mathcal{F}$ est dénombrable) suivante :

Proposition E. *On a la convergence en lois finies-dimensionnelles :*

$$\left(\sqrt{n}\,(t_{\mathrm{inj}}(F, G_n) - t(F, W)) : F \in \mathcal{F}\right) \xrightarrow[n\to\infty]{(lfd)} \Theta_{\mathrm{inj}},$$

où $\Theta_{\mathrm{inj}} = (\Theta_{\mathrm{inj}}(F) : F \in \mathcal{F})$ *est un processus gaussien centré de fonction de covariance* K_{inj} *donnée, pour* $F, F' \in \mathcal{F}$*, par :*

$$K_{\mathrm{inj}}(F, F') = \sum_{q \in V(F)} \sum_{q' \in V(F')} t\left((F \bowtie F')(q, q'), W\right) - v(F)v(F')\, t(F, W)t(F', W).$$

En particulier, on retrouve le théorème central limite donné par Féray, Méliot et Nikeghbali [84]. Ce dernier résultat nous permet aussi d'obtenir le théorème central limite pour les densités d'homomorphismes de graphes quantiques (voir la convergence (4.52)) et pour les densités d'homomorphismes induits, voir le corollaire 4.15 du chapitre 4.

2.4.3 Extensions possibles

Convergence du processus du théorème B

Dans le théorème B, nous avons établi la convergence au sens des lois finies dimensionnelles du processus $(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1))$ vers le processus $(\chi_y : y \in (0,1))$. On aimerait pouvoir montrer la convergence du processus pour la topologie de Skorokhod. Pour ce faire, un argument de tension est nécessaire. On pourrait par exemple calculer les moments croisés d'ordre 4 du processus. Les techniques développées dans le chapitre 4 devraient pouvoir fonctionner en utilisant des développements d'Edgeworth pour des sommes de vecteurs aléatoires indépendants de variables de Bernoulli corrélées. Cependant, cette approche semble très technique et il nous apparaît préférable de trouver une méthode alternative.

Résultats asymptotiques pour des FDR empiriques associées à d'autres petits graphes comme les triangles

On pourrait établir un résultat plus général sur les asymptotiques de FDR empiriques associées non plus seulement aux degrés de la suite de W-graphes aléatoires $(G_n(W) : n \in \mathbb{N}^*)$, i.e. au graphe complet à deux sommets K_2 mais à des graphes simples finis quelconques F. On pense notamment aux triangles ou aux graphes étoilés à deux arêtes (i.e. un sommet connecté par deux arêtes à deux autres sommets, « two-stars » en anglais). Pour $F \in \mathcal{F}$ tel que $V(F) = [p]$ et $\ell \in \mathcal{S}_{p,k}$, on pourrait s'intéresser au processus suivant :

$$\left(\frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \mathbf{1}_{\left\{t_{\mathrm{inj}}(F^\ell, G_n^\alpha) \le t_x(F^\ell, W)\right\}} : x \in (0,1)^k\right),$$

où $t_x(F^\ell, W) = \mathbb{E}\left[t_{\mathrm{inj}}\left(F^\ell, G_n^{[k]}\right) \middle| (X_1, \ldots, X_k) = x\right]$ d'après (4.31) et la seconde égalité de (4.37), dans le chapitre 4.

Deuxième partie

Résultats

CHAPITRE 3

FONCTIONNELLES DE COÛT DE GRANDS ARBRES ALÉATOIRES (UNIFORMES ET SIMPLEMENT GÉNÉRÉS)

Version non modifiée de l'article [54]
Cost functionals for large (uniform and simply generated) random trees
paru dans *Electronic Journal of Probability*. Des notes de bas de page ont été ajoutées pour corriger quelques inexactitudes et préciser certaines conventions.

Abstract. Additive tree functionals allow to represent the cost of many divide-and-conquer algorithms. We give an invariance principle for such tree functionals for the Catalan model (random tree uniformly distributed among the full binary ordered trees with given number of nodes) and for simply generated trees (including random tree uniformly distributed among the ordered trees with given number of nodes). In the Catalan model, this relies on the natural embedding of binary trees into the Brownian excursion and then on elementary L^2 computations. We recover results first given by Fill and Kapur (2004) and then by Fill and Janson (2009). In the simply generated case, we use convergence of conditioned Galton-Watson trees towards stable Lévy trees, which provides less precise results but leads us to conjecture a different phase transition value between "global" and "local" regimes. We also recover results first given by Janson (2003 and 2016) in the Brownian case and give a generalization to the stable case.

3.1 Introduction

Ordered trees have many applications in various fields such as computer science for data structures or in biology for genealogical or phylogenetic trees of extant species. Related to those applications, the study of large trees has attracted some attention. In this paper, we shall consider asymptotics in the global regime for general additive functionals of large trees corresponding to the Catalan model and some simply generated trees. Such additive functionals give indexes of trees which are used in computer science, physics or biology to summarize some properties of trees.

For instance, the total path length $P(\mathbf{t})$ of a tree $\mathbf{t}$, see (3.1) and (3.2) for a precise definition, which sums the distances to the root of all nodes, in the context of binary search trees, counts the number of key comparisons needed by Hoare's sorting algorithm Quicksort to sort a list of randomly permuted items, see Rösler [172]. Its convergence towards the Airy distribution was first established by Takács [183], see also Aldous [9, 10] and Janson [113] for binary trees under the Catalan model, Régnier [166], Rösler [172] for binary search trees under the random permutation model (RPM) and Fill and Kapur [90, 91] for m-ary search

trees.

The Wiener index $W(\mathbf{t})$ of a tree $\mathbf{t}$, see again definitions (3.1) and (3.2) for a precise definition, which sums the distances between all pairs of nodes of $\mathbf{t}$, was introduced by the chemist Wiener [193] in 1947. It was initially defined as the number of bonds between all pairs of atoms in an acyclic molecule. It also plays an important role in physicochemical properties of chemical structures (boiling points, heat of formation, crystal defects, ...), see Dobrynin, Entringera and Gutman [66] or Trinajstic [185], chapter 10. Its asymptotics has been studied by Janson [113] for binary trees under the Catalan model, Neininger [152] for binary search trees under the RPM and recursive trees and Janson [113] for simply generated trees.

The study of additive functionals associated with monomials, that is $f(x) = x^{\beta-1}$ in (3.1) or equivalently $b_n = n^\beta$ in (3.4), with $\beta > 0$, is interesting because many usual additive functionals can be expressed in terms of those elementary functionals. Moreover, a phase transition in the limiting behavior appears when β varies, see Fill and Kapur [89], Fill and Janson [88] for uniform binary trees, Neininger [151] for binary search trees under RPM and Fill and Kapur for m-ary trees [90, 91].

Additive functionals also appears naturally for the study of phylogenetic trees (rooted binary trees with n labeled leaves corresponding to species and $n-1$ internal vertices). When the number of species in studies of phylogenies grows, it can be interesting to look at the shapes of these trees through indices. Among these indices, we can cite the Sackin index $S(\mathbf{t})$ of a tree $\mathbf{t}$, see definition (3.7), introduced in 1972 by Sackin [175] and also studied in computer science for binary search trees (named as external path length), see Régnier [166] and Rösler [172]. Blum, François and Janson [27] studied its asymptotics. We can also consider the Colless index $C(\mathbf{t})$ of a tree $\mathbf{t}$, see definition (3.6), introduced by Colless [50] in 1982. Its asymptotics have also been studied by by Blum, François and Janson [27]. The cophenetic index $\mathrm{Co}(\mathbf{t})$ of a tree $\mathbf{t}$ was introduced in 2013 by Mir, Rosselló and Rotger [146] and Cardona, Mir and Rosselló [42] who studied its limiting behavior.

We stress that additive functionals in the local regime, such as the total size, the number of leaves, the number of protected nodes, the number of sub-trees or the shape functional (take $f(x) = \log(x)/x$ in (3.1) or $b_n = \log(n)$ in (3.4)) are not covered by our results. See Wagner [191], Holmgren et Janson [108], Janson [115] and Ralaivaosaona and Wagner [165] for asymptotic results in the local regime.

3.1.1 A finite measure indexed by a tree

Let $\mathbb{T}$ denote the set of all rooted finite ordered trees. For $\mathbf{t} \in \mathbb{T}$, let $|\mathbf{t}|$ be the the number of nodes of $\mathbf{t}$; for a node $v \in \mathbf{t}$, let $\mathbf{t}_v$ denote the sub-tree of $\mathbf{t}$ above v (see (3.11) in Section 3.2.1 for a precise definition). We consider the following unnormalized non-negative finite measure $\mathcal{A}_\mathbf{t}$:

$$\mathcal{A}_\mathbf{t}(f) = \sum_{v \in \mathbf{t}} |\mathbf{t}_v| f\left(\frac{|\mathbf{t}_v|}{|\mathbf{t}|}\right), \tag{3.1}$$

where f is a measurable real-valued function defined on $[0,1]$. We are interested in the asymptotic distribution of $\mathcal{A}_\mathbf{t}(f)$ when $\mathbf{t}$ belongs to a certain class of trees and $|\mathbf{t}|$ goes to infinity. We shall consider two classes of trees: the binary trees (and more precisely the Catalan model) and some simply generated trees.

We give some examples related to the measure $\mathcal{A}_\mathbf{t}$ which are commonly used in the analysis of trees. In what follows, for a tree $\mathbf{t} \in \mathbb{T}$, we denote by $\emptyset$ its root and by d the usual graph distance on $\mathbf{t}$. For $v, w \in \mathbf{t}$, we say that w is an ancestor of v and write $w \preccurlyeq v$ if $d(\emptyset, v) = d(\emptyset, w) + d(w, v)$. For $u, v \in \mathbf{t}$, we denote by $u \wedge v$, the most recent common ancestor of u and v: $u \wedge v$ is the only element of $\mathbf{t}$ such that: $w \preccurlyeq u$ and $w \preccurlyeq v$ implies $w \preccurlyeq u \wedge v$.

- The **total path length** of $\mathbf{t}$ is defined by $P(\mathbf{t}) = \sum_{w \in \mathbf{t}} d(\emptyset, w)$. As $d(\emptyset, w) = \sum_{v \in \mathbf{t}} \mathbf{1}_{\{v \preccurlyeq w\}} -$

1, we get: $P(\mathbf{t}) = \sum_{v\in\mathbf{t}} \sum_{w\in\mathbf{t}} \mathbf{1}_{\{v\preccurlyeq w\}} - |\mathbf{t}| = \mathcal{A}_{\mathbf{t}}(1) - |\mathbf{t}|$ [1].

- The **shape functional** of $\mathbf{t}$ is defined by $\sum_{w\in\mathbf{t}} \log(|\mathbf{t}_w|)$. We also have the equality $\sum_{w\in\mathbf{t}} \log(|\mathbf{t}_w|) = |\mathbf{t}|^{-1}\mathcal{A}_{\mathbf{t}}(\log(x)/x) + |\mathbf{t}|\log(|\mathbf{t}|)$. (The function $\log(x)/x$ will not be covered by the main results of this paper.)

- The **Wiener index** of $\mathbf{t}$ is defined by $W(\mathbf{t}) = \sum_{u,w\in\mathbf{t}} d(u,w)$. Since

$$d(u,w) = \sum_{v\in\mathbf{t}} (\mathbf{1}_{\{v\preccurlyeq u\}} + \mathbf{1}_{\{v\preccurlyeq w\}} - 2\mathbf{1}_{\{v\preccurlyeq u,\, v\preccurlyeq w\}}),$$

we deduce that $W(\mathbf{t}) = 2|\mathbf{t}|\,(\mathcal{A}_{\mathbf{t}}(1) - \mathcal{A}_{\mathbf{t}}(x))$.

In a nutshell, for $\mathbf{t}\in\mathbb{T}$, we have:

$$\Big(P(\mathbf{t}),\, W(\mathbf{t})\Big) = \Big(\mathcal{A}_{\mathbf{t}}(1) - |\mathbf{t}|,\, 2|\mathbf{t}|(\mathcal{A}_{\mathbf{t}}(1) - \mathcal{A}_{\mathbf{t}}(x))\Big). \tag{3.2}$$

The measure $\mathcal{A}_{\mathbf{t}}$ is also related to other additive functionals in the particular case of binary trees, see Section 3.1.2.

3.1.2 Additive functionals and toll functions for binary trees

Additive functionals on binary trees allow to represent the cost of algorithms such as "divide and conquer", see Fill and Kapur [89]. For $\mathbf{t}\in\mathbb{T}$ a full binary tree, we shall denote by 1 (resp. 2) the left (resp. right) child of the root. Thus $\mathbf{t}_1$ (resp. $\mathbf{t}_2$) will be the left (resp. right) sub-tree of the root of $\mathbf{t}$. A functional F on binary trees is called an additive functional if it satisfies the following recurrence relation:

$$F(\mathbf{t}) = F(\mathbf{t}_1) + F(\mathbf{t}_2) + b_{|\mathbf{t}|}, \tag{3.3}$$

for all trees $\mathbf{t}$ such that $|\mathbf{t}|\geq 3$ and with $F(\{\emptyset\}) = b_1$. The given sequence $(b_n, n\in\mathbb{N}^*)$ is called the toll function. Notice that:

$$F(\mathbf{t}) = \sum_{v\in\mathbf{t}} b_{|\mathbf{t}_v|}. \tag{3.4}$$

In the particular case where the toll function is a power function, that is $b_n = n^\beta$ for $n\in\mathbb{N}^*$ and some $\beta > 0$, we get $F(\mathbf{t}) = |\mathbf{t}|^{-\beta+1}\mathcal{A}_{\mathbf{t}}(x^{\beta-1})$. In such cases, the asymptotic study of the measure $\mathcal{A}_{\mathbf{t}}$ will provide the asymptotic of the additive functionals.

We say that $v\in\mathbf{t}$ is a leaf if $|\mathbf{t}_v| = 1$. We denote by $\mathcal{L}(\mathbf{t})$ the set of leaves of $\mathbf{t}$ and, when $|\mathbf{t}| > 1$, by $\mathbf{t}^* = \mathbf{t}\setminus\mathcal{L}(\mathbf{t})$ the tree $\mathbf{t}$ without its leaves. We stress that the additive functional considered in [89] is exactly

$$\tilde{F}(\mathbf{t}) = F(\mathbf{t}^*) = \sum_{v\in\mathbf{t}^*} b_{|\mathbf{t}^*_v|}. \tag{3.5}$$

However the asymptotics will be the same as the one for F when the toll function is a power function, see Remark 3.9. We complete the examples of the previous section for binary trees.

- The **Sackin index** (or external path length) of a tree $\mathbf{t}$, used to study the balance of the tree, is similar to the total path length of $\mathbf{t}$ when one considers only the leaves: $S(\mathbf{t}) = \sum_{w\in\mathcal{L}(\mathbf{t})} d(\emptyset, w)$. Using that for a full binary tree we have $|\mathbf{t}| = 2|\mathcal{L}(\mathbf{t})| - 1$, we deduce that $2S(\mathbf{t}) = \sum_{v\in\mathbf{t}} |\mathbf{t}_v| - 1 = \mathcal{A}_{\mathbf{t}}(1) - 1$.

[1] By convention, for $a\in\mathbb{R}$, we denote the function $x\mapsto x^a\mathbf{1}_{(0,1]}(x)$ defined on $[0,1]$ by x^a.

- The **Colless index** of a binary tree $\mathbf{t}$ is defined as $C(\mathbf{t}) = \sum_{v\in\mathbf{t}^*} |L_v - R_v|$, where $L_v = |\mathcal{L}(\mathbf{t}_{v1})|$ (resp. $R_v = |\mathcal{L}(\mathbf{t}_{v2})|$) is the number of leaves of the left (resp. right) sub-tree above v. Since $\mathbf{t}$ is a full binary tree, we get $2L_v - 2R_v = |\mathbf{t}_{v1}| - |\mathbf{t}_{v2}|$ and $|\mathbf{t}_{v1}| + |\mathbf{t}_{v2}| = |\mathbf{t}_v| - 1$. We obtain that $2C(\mathbf{t}) = \sum_{v\in\mathbf{t}} |\mathbf{t}_v| - |\mathbf{t}| - 2\chi(\mathbf{t})$ [2], with

$$\chi(\mathbf{t}) = \sum_{v\in\mathbf{t}^*} \min(|\mathbf{t}_{v1}|, |\mathbf{t}_{v2}|). \tag{3.6}$$

  That is $2C(\mathbf{t}) = \mathcal{A}_{\mathbf{t}}(1) - |\mathbf{t}| - 2\chi(\mathbf{t})$.

- The **cophenetic index** of a tree $\mathbf{t}$ (which is used in [146] to study the balance of the tree) is defined by $\mathrm{Co}(\mathbf{t}) = \sum_{u,w\in\mathcal{L}(\mathbf{t}),\, u\neq w} d(\emptyset, u\wedge w)$. Using again that $\mathbf{t}$ is a full binary tree, we get $4\mathrm{Co}(\mathbf{t}) = 4\sum_{v\in\mathbf{t}} |\mathcal{L}(\mathbf{t}_v)|(|\mathcal{L}(\mathbf{t}_v)| - 1) - 4|\mathcal{L}(\mathbf{t})|(|\mathcal{L}(\mathbf{t})| - 1) = \sum_{v\in\mathbf{t}} |\mathbf{t}_v|^2 - |\mathbf{t}|^2 - |\mathbf{t}| + 1$. That is $4\mathrm{Co}(\mathbf{t}) = |\mathbf{t}|\mathcal{A}_{\mathbf{t}}(x) - |\mathbf{t}|^2 - |\mathbf{t}| + 1$.

In a nutshell, for $\mathbf{t} \in \mathbb{T}$ full binary, we have:

$$\Big(2S(\mathbf{t}), 2C(\mathbf{t}), 4\mathrm{Co}(\mathbf{t})\Big) = \Big(\mathcal{A}_{\mathbf{t}}(1) - 1,\ \mathcal{A}_{\mathbf{t}}(1) - |\mathbf{t}| - 2\chi(\mathbf{t}),\ |\mathbf{t}|\mathcal{A}_{\mathbf{t}}(x) - |\mathbf{t}|^2 - |\mathbf{t}| + 1\Big). \tag{3.7}$$

3.1.3 Main results on the asymptotics of additive functionals in the Catalan model

We consider the Catalan model: let T_n be a random tree uniformly distributed among the set of full binary ordered trees with n internal nodes (and thus $n+1$ leaves), which has cardinality $C_n = (2n)!/[(n!^2)(n+1)]$. We have:

$$\boxed{|\mathrm{T}_n| = 2n + 1}.$$

Recall that T_n is a (full binary) Galton-Watson tree (also known as simply generated tree) conditioned on having n internal nodes (see Janson [114], Example 10.3). It is well known, see Takàcs [183], Aldous [9, 10] and Janson [113], that $|\mathrm{T}_n|^{-3/2}P(\mathrm{T}_n)$ converges in distribution, as n goes to infinity, towards $2\int_0^1 B_s\, ds$, where $B = (B_s, s \in [0,1])$ is the normalized positive Brownian excursion. This result, see Corollary 3.16, can be seen as a consequence of the convergence in distribution of T_n (in fact the contour process) properly scaled towards the Brownian continuum tree whose contour process is B, see [9] and Duquesne [69], or Duquesne and Le Gall [70] in the setting of Brownian excursion. For a combinatorial approach, which can be extended to other families of trees, see also Fill and Kapur [90, 91] or Fill, Flajolet and Kapur [86].

In [89], the authors considered the toll functions $b_n = n^\beta$ with $\beta > 0$ and they proved that with a suitable scaling the corresponding additive functional $F_\beta(\mathrm{T}_n) = |\mathrm{T}_n|^{-\beta+1}\mathcal{A}_{\mathrm{T}_n}(x^{\beta-1})$ converge in distribution to a limit, say Y_β. The distribution of Y_β is characterized by its moments. (In [85, 89], the authors considered also the toll function $b_n = \log(n)$.) See also Janson and Chassaing [116] for asymptotics of the Wiener index, which is a consequence of the joint convergence in distribution of $(\mathcal{A}_{\mathrm{T}_n}(1), \mathcal{A}_{\mathrm{T}_n}(x))$ with a suitable scaling and Blum, François and Janson [27] for the convergence of the Sackin and Colless indexes. We give a natural representation of the family $(\mathrm{T}_n, n \in \mathbb{N}^*)$ such that we have an a.s. convergence of the additive functional instead of a convergence in distribution (see Section 3.2.4). In Theorem 3.4 (take $\alpha = 2$), we prove that, in the Catalan model, the random measure $|\mathrm{T}_n|^{-3/2}\mathcal{A}_{\mathrm{T}_n}$ converges weakly a.s., as n goes to infinity, to a random measure $2\Phi_B$, built on the Brownian normalized excursion B, see (3.16) with $h = B$. Using the notation $\mathrm{T}_{n,v} = (\mathrm{T}_n)_v$ for $v \in \mathrm{T}_n$, this proves in particular the following a.s. convergence. See also the fluctuations for this a.s. convergence, in Proposition 3.10.

[2] By convention, $\sum_k a_k + b = (\sum_k a_k) + b$.

Theorem. *We have that a.s. for all real-valued continuous function f defined on $[0,1]$:*

$$|\mathrm{T}_n|^{-3/2} \sum_{v\in \mathrm{T}_n} |\mathrm{T}_{n,v}|\; f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) \xrightarrow[n\to+\infty]{a.s.} 2\Phi_B(f). \tag{3.8}$$

Notice that Theorem 3.4 is more general as the convergences hold jointly for all measurable real-valued functions f defined on $[0,1]$ such that f is continuous on $(0,1]$ and $\sup_{x\in(0,1]} x^a|f(x)|$ is finite for some $a < 1/2$. Notice this covers the case of toll functions $b_n = n^\beta$ with $\beta > 1/2$ in [89] which corresponds to the so called "global" regime. The limit $2\Phi_B(x^{\beta-1})$ gives a representation of Y_β for $\beta > 1/2$, which, thanks to Corollary 3.6, corresponds when $\beta \geq 1$ to the one announced in Fill and Janson [88]. In particular, we have the following representation for Φ_B on monomials, see Lemma 3.5:

Corollary. *We have, for all $\beta > 1$:*

$$\Phi_B(x^{\beta-1}) = \frac{1}{2}\,\beta(\beta-1)\int_{[0,1]^2} |t-s|^{\beta-2}\, m_B(s,t)\, ds\, dt,$$

where $m_B(s,t) = \inf_{u\in[s\wedge t, s\vee t]} B(u)$.

In the "local" regime, that is $\beta \in (0,1/2]$, according to Corollary 3.6 and Lemma 3.5, the convergence (3.8) is not relevant as $\Phi_B(x^{\beta-1}) = +\infty$ a.s.; see [89] for the relevant normalization.

The proof of Theorem 3.4 relies on the natural embedding of T_n into the Brownian excursion, see [10] and Le Gall [134], so that the convergence in distribution of the random measure $|\mathrm{T}_n|^{-3/2}\mathcal{A}_{\mathrm{T}_n}$ or of the additive functionals F_β (which holds simultaneously for all $\beta > 1/2$) is then an a.s. convergence. In Remark 3.8, we provide, as a direct consequence of Theorem 3.4, the joint convergence of the total length path, the Wiener, Sackin, Colless and cophenetic indexes defined in Sections 3.1.1 and 3.1.2.

Remark 3.1. The method presented in this section based on the embedding of T_n into a Brownian excursion can not be extended directly to other models of trees such as binary search trees, recursive trees or simply generated trees.

Concerning binary search trees (or random permutation model or Yule trees), see [166] and [172] for the convergence of the external path length (which corresponds in our setting to the Sackin index), [151] for toll function $b_n = n^\beta$, [152] for the Wiener index (and [113] for simply generated trees), [27] (and [93] for other trees) for the Sackin and Colless indexes, and [85] for the shape function.

Concerning recursive trees, see [141, 65] for the convergence of the total path length and [152] for the Wiener index. In the setting of recursive trees, then (3.3) is a stochastic fixed point equation, which can be analyzed using the approach of [173].

Remark 3.2. One can replace the toll function $b_{|\mathbf{t}|}$ in (3.3) by a function of the tree, say $\mathbf{b}(\mathbf{t})$. For example, if one consider $\mathbf{b}(\mathbf{t}) = \mathbf{1}_{\{\mathbf{t}=\mathbf{t}_0\}}$, with $\mathbf{t}_0$ a given tree, then the corresponding additive functional gives the number of occurrence of the motif $\mathbf{t}_0$. The case of "local" toll function $\mathbf{b}$ (with finite support or fast decreasing rate) has been considered in the study of fringe trees, see [7], [60, 92] for binary search trees, and [115] for simply generated trees and [108] for binary search trees and recursive trees.

The terms "local" and "global" are used to stress the phase transition of the limit laws from normal to non-normal. If the toll function is small then the contribution $b_{|\mathbf{t}|}$ from each sub-tree $\mathbf{t}$ is small so that the limit law is normal. But, if the toll function is large, then the main contribution comes from a few sub-trees of large size so that the limit law is non-normal. See [111] for the study of the phase transition on asymptotics of additive functionals with toll functions $b_n = n^\beta$ on binary search trees between the "local" regime (corresponding to $\beta \leq 1/2$) and the "global" regime ($\beta > 1/2$). The same phase transition is observed for the Catalan model, see [89]. Our main result, see Theorem 3.4, concerns specifically the "global" regime.

3.1.4 Main results on the asymptotics of additive functionals for simply generated trees

We consider a weight sequence $\mathfrak{p} = (\mathfrak{p}(k), k \in \mathbb{N})$ on $\mathbb{R}_+$ with generating function $g_\mathfrak{p}$. We assume that $g_\mathfrak{p}$ has a positive radius of convergence, $g_\mathfrak{p}(0) > 0$, $g''_\mathfrak{p} \neq 0$ and $\mathfrak{p}$ is generic, that is there exists a positive root to the equation $g_\mathfrak{p}(q) = qg'_\mathfrak{p}(q)$. A simply generated tree of size $p \in \mathbb{N}^*$ with weight function $\mathfrak{p}$ is a random tree $\tau^{(p)}$ such that the probability of $\tau^{(p)}$ to be equal to $\mathbf{t}$, with $|\mathbf{t}| = p$, is proportional to $\prod_{v\in\mathbf{t}} \mathfrak{p}(k_v(\mathbf{t}))$, where $k_v(\mathbf{t})$ is the number of children of the node v in $\mathbf{t}$. According to Section 3.2.5, since $g_\mathfrak{p}$ is generic, without loss of generality we can assume that $\mathfrak{p}$ is a critical probability ($g_\mathfrak{p}(1) = g'_\mathfrak{p}(1) = 1$), so that $\tau^{(p)}$ is distributed as a Galton-Watson (GW) tree τ with offspring distribution $\mathfrak{p}$ conditioned to $|\tau| = p$. Global convergence of scaled GW trees τ to Lévy trees has been studied in Le Gall and Le Jan [136] and in [70] using the convergence of contour process.

Assume $\mathfrak{p}$ belongs to the domain of attraction of a stable distribution of Laplace exponent $\psi(\lambda) = \kappa\lambda^\gamma$ with $\gamma \in (1, 2]$ and $\kappa > 0$. Then, the convergence of $\tau^{(p)}$ properly scaled to the normalized Lévy trees holds according to [69]. This result is recalled in Section 3.4.2. We recall that the normalized Lévy tree is a real tree coded by the normalized positive excursion of the height function $H = (H(s), s \in [0, 1])$.

We have the following convergence in distribution, see Corollary 3.15 for a precise statement.

Theorem. *There exists a sequence $(a_p, p \in \mathbb{N}^*$ s.t. $\mathbb{P}(|\tau| = p) > 0)$ such that we have the following convergence in distribution:*

$$\frac{a_p}{p^2} \sum_{v\in\tau^{(p)}} |\tau_v^{(p)}|\, f\left(\frac{|\tau_v^{(p)}|}{p}\right) \xrightarrow[p\to+\infty]{(d)} \Phi_H(f), \tag{3.9}$$

simultaneously for all real-valued continuous function f defined on $[0, 1]$ [3].

The convergence (3.9) has to be understood along the infinite sub-sequence of p such that $\mathbb{P}(|\tau| = p) > 0$. The proof relies on the fact that one can approximate $\mathcal{A}_\mathbf{t}(x^k)$, for $k \in \mathbb{N}^*$, by an elementary continuous functional of the contour process of $\mathbf{t}$, see Section 3.4.4. Then, we use the convergence of the contour process of $\tau^{(p)}$ to the contour process of H to conclude. We also provide the first moment of $\Phi_H(x^{\beta-1})$, see Lemma 3.17 and conjecture that $\beta = 1/\gamma$ corresponds to the phase transition between the "global" and "local" regime in this setting.

Remark 3.3. We make the following comments.

- Assume that $\mathfrak{p}$ has finite variance, say σ^2. Then one can take $a_p = \sqrt{p}$ and H is equal to $(2/\sigma)B$ which corresponds to $\psi(\lambda) = \sigma^2\lambda^2/2$. By scaling, or using that the limit in Theorem 3.4 does not depend on α, we deduce that $\Phi_{cB} = c\Phi_B$. We can then rewrite (3.9) as:

$$p^{-3/2} \sum_{v\in\tau^{(p)}} |\tau_v^{(p)}|\, f\left(\frac{|\tau_v^{(p)}|}{p}\right) \xrightarrow[p\to+\infty]{(\mathrm{d})} \frac{2}{\sigma}\Phi_B(f), \tag{3.10}$$

  where the convergence holds simultaneously for all real-valued continuous function f defined on $[0, 1]$ and along the infinite sub-sequence of p such that $\mathbb{P}(|\tau| = p) > 0$.

- If one consider the binary offspring distribution $\mathfrak{p}$ such that $\mathfrak{p}(2) + \mathfrak{p}(0) = 1$ (recall that $1 > \mathfrak{p}(0) > 0$ by assumption), one gets that $\tau^{(2n+1)}$ is uniformly distributed among the full binary trees with n internal nodes (and $n+1$ leaves), that is $\tau^{(2n+1)}$ is distributed as T_n, see the Catalan model studied in Section 3.1.3. Take $\mathfrak{p}(0) = 1/2$ to get the critical

[3] The right wording is in terms of the convergence of measures given in Corollary 3.15.

case, and notice that $\sigma = 1$ in (3.10). The convergence (3.10), with $p = 2n + 1$, is then a weaker version of (3.8) (convergence in distribution instead of a.s. convergence, and continuous functions on $[0, 1]$ instead of continuous functions on $(0, 1]$ with possible blow up at 0+).

- If one consider the (shifted) geometric distribution: $\mathfrak{p}(k) = q(1-q)^k$ for $k \in \mathbb{N}$ with $q \in (0, 1)$, one gets that $\tau^{(p)}$ is uniformly distributed among the rooted ordered trees with p nodes. Take $q = 1/2$ to get the critical case, and notice that $\sigma = 2$ in (3.10).

- If one consider the Poisson offspring distribution: $\mathfrak{p}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for $k \in \mathbb{N}$ with $\lambda \in \mathbb{R}^+$, one gets that $\tau^{(p)}$ is uniformly distributed among the labeled unordered rooted trees with p nodes (also known as Cayley trees). Take $\lambda = 1$ to get the critical case, and notice that $\sigma = 1$ in (3.10). In particular, we recover the result of Zohoorian Azad [196] for the additive functional associated to the toll function $b_n = n^2$ for $n \in \mathbb{N}^*$ (apply Corollary 3.16 with $\beta = 2$).

3.1.5 Organization of the paper

Section 3.2 is devoted to the definition of the main objects used in this paper (ordered rooted discrete trees using Neveu's formalism, real trees defined by a contour function, Brownian tree whose contour function is a Brownian normalized excursion, the embedding of the discrete binary trees from the Catalan model into the Brownian tree, and simply generated random trees). We present our main result about the Catalan model in Section 3.3 on the convergence (3.8), see Theorem 3.4 and Corollary 3.6. (The proofs are given in Sections 3.5 and 3.6.) The corresponding fluctuations are stated in Proposition 3.10. (The proof is given in Section 3.7.) Section 3.4 is devoted to the main results concerning the convergence of $\mathcal{A}_\tau$ when τ is a simply generated tree, see Corollaries 3.15 and 3.16. Some technical results are gathered in Section 3.8.

3.2 Notations

Let I be an interval of $\mathbb{R}$ with positive Lebesgue measure. We denote by $\mathcal{B}(I)$ the set of real-valued measurable functions defined on I. We denote by $\mathcal{C}(I)$ (resp. $\mathcal{C}_+(I)$) the set of real-valued (resp. non-negative) continuous functions defined on I. For $f \in \mathcal{B}(I)$ we denote by $\| f \|_\infty$ the supremum norm and by $\| f \|_{\text{esssup}}$ the essential supremum of $|f|$ over I. The two supremums coincide when f is continuous.

3.2.1 Ordered rooted discrete trees

We recall Neveu's formalism [153] for ordered rooted discrete trees, which we shall simply call trees. We set $\mathcal{U} = \bigcup_{n \geq 0} (\mathbb{N}^*)^n$ the set of finite sequences of positive integers with the convention $(\mathbb{N}^*)^0 = \{\emptyset\}$. For $n \geq 0$ and $u \in (\mathbb{N}^*)^n \subset \mathcal{U}$, we set $|u| = n$ the length of u. Let $u, v \in \mathcal{U}$. We denote by uv the concatenation of the two sequences, with the convention that $uv = u$ if $v = \emptyset$ and $uv = v$ if $u = \emptyset$. We say that v is an ancestor of u (in a large sense) and write $v \preccurlyeq u$ if there exists $w \in \mathcal{U}$ such that $u = vw$. If $v \preccurlyeq u$ and $v \neq u$, then we shall write $v \prec u$. The set of ancestors of u is the set $\bar{\mathrm{A}}_u = \{v \in \mathcal{U}; v \preccurlyeq u\}$. The most recent common ancestor of a subset $\mathbf{s}$ of $\mathcal{U}$, denoted by $\mathfrak{m}(\mathbf{s})$, is the unique element v of $\bigcap_{u \in \mathbf{s}} \bar{\mathrm{A}}_u$ with maximal length. We consider the lexicographic order on $\mathcal{U}$: for $u, v \in \mathcal{U}$, we set $v < u$ either if $v \prec u$ or if $v = wjv'$ and $u = wiu'$ with $w = \mathfrak{m}(\{v, u\})$, $u, u' \in \mathcal{U}$ and $j < i$ for some $i, j \in \mathbb{N}^*$.

A tree $\mathbf{t}$ is a subset of $\mathcal{U}$ that satisfies:

- $\emptyset \in \mathbf{t}$,
- If $u \in \mathbf{t}$, then $\bar{\mathrm{A}}_u \subset \mathbf{t}$.
- For every $u \in \mathbf{t}$, there exists $k_u(\mathbf{t}) \in \mathbb{N}$ such that, for every $i \in \mathbb{N}^*$, $ui \in \mathbf{t}$ if and only if $1 \leq i \leq k_u(\mathbf{t})$.

Let $u \in \mathbf{t}$. The integer $k_u(\mathbf{t})$ represents the number of offsprings of the node u. The node u is called a leaf (resp. internal node) if $k_u(\mathbf{t}) = 0$ (resp. $k_u(\mathbf{t}) > 0$). The node $\emptyset$ is called the root of $\mathbf{t}$. We define the sub-tree $\mathbf{t}_u \in \mathbb{T}$ of $\mathbf{t}$ "above" u as:

$$\mathbf{t}_u = \{v \in \mathcal{U},\ uv \in \mathbf{t}\}. \tag{3.11}$$

We denote by $|\mathbf{t}| = \mathrm{Card}\,(\mathbf{t})$ the number of nodes of $\mathbf{t}$ and we say that $\mathbf{t}$ is finite if $|\mathbf{t}| < +\infty$. Let $d_{\mathbf{t}}$ denote the usual graph distance on $\mathbf{t}$. In particular, we have $d_{\mathbf{t}}(\emptyset, u) = |u|$ for $u \in \mathbf{t}$. When the context is clear, we shall write d for $d_{\mathbf{t}}$.

We denote by $\mathbb{T}$ the set of finite trees and by $\mathbb{T}^{(p)} = \{\mathbf{t} \in \mathbb{T},\ |\mathbf{t}| = p\}$ the set of trees with p nodes, for $p \in \mathbb{N}^*$. Let us recall that, for a tree $\mathbf{t} \in \mathbb{T}$, we have

$$\sum_{u \in \mathbf{t}} k_u(\mathbf{t}) = |\mathbf{t}| - 1. \tag{3.12}$$

3.2.2 Real trees

We recall the definition of a real tree, see [78]. A real tree is a metric space $(\mathcal{T}, d)$ which satisfies the following two properties for every $x, y \in \mathcal{T}$:

(i) There exists a unique isometric map $f_{x,y}$ from $[0, d(x,y)]$ into $\mathcal{T}$ such that $f_{x,y}(0) = x$ and $f_{x,y}(d(x,y)) = y$.

(ii) If ϕ is a continuous injective map from $[0,1]$ into $\mathcal{T}$ such that $\phi(0) = x$ and $\phi(1) = y$, then we have $\phi([0,1]) = f_{x,y}([0, d(x,y)])$.

Equivalently, a metric space $(\mathcal{T}, d)$ is a real tree if and only if $\mathcal{T}$ is connected and d satisfies the four point condition:

$$d(s,t) + d(x,y) \leq \max(d(s,x) + d(t,y), d(s,y) + d(t,x)) \quad \text{for all} \quad s,t,x,y \in \mathcal{T}.$$

A rooted real tree is a real tree $(\mathcal{T}, d)$ with a distinguished element $\emptyset$ called the root. For $x, y \in \mathcal{T}$, we denote by $[\![x, y]\!]$ the range of the map $f_{x,y}$ described above. Let $x, y \in \mathcal{T}$. We denote by $x \wedge y$ their most recent common ancestor which is the only $z \in \mathcal{T}$ such that $[\![\emptyset, z]\!] = [\![\emptyset, x]\!] \bigcap [\![\emptyset, y]\!]$. The out-degree $d_x(\mathcal{T})$ of x is the number of connected components of $\mathcal{T} \backslash \{x\}$ which do not contain the root. We say x is a leaf (resp. branching point) if $d_x(\mathcal{T}) = 0$ (resp. $d_x(\mathcal{T}) \geq 2$). We say $\mathcal{T}$ is binary if $d_x(\mathcal{T}) \in \{0, 1, 2\}$ for all $x \in \mathcal{T}$.

For $h \in \mathcal{C}_+([0,1])$, we define its minimum over the interval with bounds $s, t \in [0,1]$:

$$m_h(s,t) = \inf_{u \in [s \wedge t, s \vee t]} h(u). \tag{3.13}$$

We shall also use the length of the excursion of h above level r straddling s defined by:

$$\sigma_{r,s}(h) = \int_0^1 dt\, \mathbf{1}_{\{m_h(s,t) \geq r\}}. \tag{3.14}$$

For $\beta > 0$, we set:

$$Z_\beta^h = \int_0^1 ds \int_0^{h(s)} dr\ \sigma_{r,s}(h)^{\beta - 1}. \tag{3.15}$$

Let $h \in \mathcal{C}_+([0,1])$ be such that $m_h(0,1) = 0$. For every $x, y \in [0,1]$, we set $d_h(x,y) = h(x) + h(y) - 2m_h(x,y)$. It is easy to check that d_h is symmetric and satisfies the triangle inequality. The relation $\sim_h$ defined on $[0,1]^2$ by $x \sim_h y \Leftrightarrow d_h(x,y) = 0$ is an equivalence relation. Let $\mathcal{T}_h = [0,1]/\sim_h$ be the corresponding quotient space. The function d_h on $[0,1]^2$ induces a function on $\mathcal{T}_h^2$, which we still denoted by d_h, and which is a distance on $\mathcal{T}_h$. It is not difficult to check that $(\mathcal{T}_h, d_h)$ is then a compact real tree. We denote by $\mathbf{p}_h$ the canonical projection from $[0,1]$ into $\mathcal{T}_h$. Thus, the metric space $(\mathcal{T}_h, d_h)$ can be viewed as a rooted real tree by setting $\emptyset = \mathbf{p}_h(0)$. The image of the Lebesgue measure on $[0,1]$ by $\mathbf{p}_h$ is a measure μ_h on $\mathcal{T}_h$.

For $\mathbf{t} \in \mathbb{T}$, we define the unnormalized measure $\mathcal{A}_\mathbf{t}$ on $[0,1]$ by:

$$\mathcal{A}_\mathbf{t}(f) = \sum_{v \in \mathbf{t}} |\mathbf{t}_v| f\left(\frac{|\mathbf{t}_v|}{|\mathbf{t}|}\right), \quad f \in \mathcal{C}([0,1]).$$

For $h \in \mathcal{C}_+([0,1])$, we also consider the measure Φ_h on $[0,1]$ defined by:

$$\Phi_h(f) = \int_0^1 ds \int_0^{h(s)} dr\ f(\sigma_{r,s}(h)), \quad f \in \mathcal{B}([0,1]). \tag{3.16}$$

We endow the space of non-negative finite measures on $[0,1]$ with the topology of weak convergence.

3.2.3 The Brownian continuum random tree $\mathcal{T}$

Let $B = (B_t, 0 \le t \le 1)$ be a positive normalized Brownian excursion. Informally, B is just a linear standard Brownian path started from the origin and conditioned to stay positive on $(0,1)$ and to come back to 0 at time 1. For $\alpha > 0$, let $e = \sqrt{2/\alpha}\, B$ and let $\mathcal{T}_e$ denote the associated real tree called Brownian continuum random tree. (We recall the associated branching mechanism is $\psi(\lambda) = \alpha\lambda^2$.) The continuum random tree introduced in [8] corresponds to $\alpha = 1/2$ and the Brownian tree associated to the normalized Brownian excursion corresponds to $\alpha = 2$. We shall keep the parameter α so that the two previous cases are easy to read on the results. See [135] for properties of the Brownian continuum random tree. In particular $\mu_e(dx)$-a.s. x is a leaf and a.s. $\mathcal{T}_e$ is binary.

We shall forget to stress the dependence in e in the notations, when there is no ambiguity, so that for example we simply write $\mathcal{T}$, μ, $\sigma_{r,s}$ and Z_β for respectively $\mathcal{T}_e$, μ_e, $\sigma_{r,s}(e)$ which is defined in (3.14) and Z_β^e which is defined in (3.15). For $r \ge 0$ and $s \in [0,1]$, we also have:

$$\sigma_{r,s} = \mu(x \in \mathcal{T},\ d(\emptyset, x \wedge \mathbf{p}(s)) \ge r)),$$

which is the mass of the sub-tree of $\mathcal{T}$ containing $\mathbf{p}(s)$ and at distance r from the root.

3.2.4 The discrete binary tree from the Brownian tree

A marked tree $\tilde{\mathbf{t}} = (\mathbf{t}, (h_v, v \in \mathbf{t}))$ is a tree $\mathbf{t} \in \mathbb{T}$ with a label on each node. The label $h_v \in (0, +\infty)$ will be interpreted as the length of the branch from below v. (Notice, there is a branch below the root.) We define the concatenation of two marked trees $\tilde{\mathbf{t}}^{(i)} = (\mathbf{t}^{(i)}, (h_v^{(i)}, v \in \mathbf{t}^{(i)}))$ with $i \in \{1,2\}$ and $r > 0$ as $\tilde{\mathbf{t}} = [\tilde{\mathbf{t}}^{(1)}, \tilde{\mathbf{t}}^{(2)}; r]$ with $\mathbf{t} = \{\emptyset\} \bigcup_{i=1}^2 \{v = iu,\ u \in \mathbf{t}^{(i)}\}$ and for $v \in \mathbf{t}$, we have $h_v = r$ if $v = \emptyset$ and $h_v = h_u^{(i)}$ if $v = iu$ with $u \in \mathbf{t}^{(i)}$ and $i \in \{1,2\}$.

Let $g \in \mathcal{C}_+([0,1])$ be such that $\mathcal{T}_g$ is binary. Let $n \in \mathbb{N}$ and $0 < t_1 < \cdots < t_{n+1} < 1$ such that $(\mathbf{p}_g(t_k), 1 \le k \le n+1)$ are $n+1$ distinct leaves. Set $G_n = (g; t_1, \ldots, t_{n+1})$. We denote by $\mathcal{T}_g(G_n) = \bigcup_{k=1}^{n+1} [\![\emptyset, \mathbf{p}_g(t_k)]\!]$ the random real tree spanned by the $n+1$ leaves $\mathbf{p}_g(t_1), \ldots, \mathbf{p}_g(t_{n+1})$ with root $\emptyset$. We define recursively the associated marked tree $\tilde{\mathbf{t}}(G_n) =$

$(\mathbf{t}(G_n), (h_{n,v}(G_n), v \in \mathbf{t}(G_n)))$, where intuitively $\mathbf{t}(G_n)$ is similar to $\mathcal{T}_g(G_n)$ but with the branch lengths equal to 1 and no branch below the root, and $h_{n,v}(G_n)$ is the length of the branch in $\mathcal{T}_g(G_n)$ below the node corresponding to $v \in \mathbf{t}(G_n)$. More precisely, for $n = 0$, we set $\mathbf{t}(G_0) = \{\emptyset\}$ and $h_{0,\emptyset}(G_0) = g(t_1)$. Let $n \geq 1$. Since $\mathcal{T}_g$ is binary and $(\mathbf{p}_g(t_k), 1 \leq k \leq n+1)$ are $n+1$ distinct leaves, there exists a unique $s \in (t_1, t_{n+1})$ and a unique $\ell \in \{1, \ldots, n\}$ such that $g(s) = m_g(t_1, t_{n+1})$ and $t_\ell < s < t_{\ell+1}$. We define $g_1(t) = (g(t) - g(s))\mathbf{1}_{[t_1,s]}(t)$ and $g_2(t) = (g(t) - g(s))\mathbf{1}_{[s,t_{n+1}]}(t)$. Notice that $\mathcal{T}_{g_i}$ is binary and $(\mathbf{p}_g(t_k), 1 \leq k \leq \ell)$ (resp. $(\mathbf{p}_g(t_k), \ell + 1 \leq k \leq n+1)$) are ℓ (resp. $n - \ell + 1$) distinct leaves of $\mathcal{T}_{g_1}$ (resp. $\mathcal{T}_{g_2}$). Set $G'_{\ell-1} = (g_1; t_1, \ldots, t_\ell)$ and $G''_{n-\ell} = (g_2; t_{\ell+1}, \ldots, t_{n+1})$ and define $\tilde{\mathbf{t}}(G_n)$ as the concatenation $[\mathbf{t}(G'_{\ell-1}), \mathbf{t}(G''_{n-\ell}); g(s)]$.

Let e be the Brownian excursion defined in Section 3.2.3. Let $(U_n, n \in \mathbb{N}^*)$ be a sequence of independent random variables uniform on $[0,1]$, independent of e. In particular $(\mathbf{p}(U_n), n \in \mathbb{N}^*)$ are a.s. distinct leaves of $\mathcal{T}$. Let $(U_{1,n}, \ldots, U_{n+1,n})$ be the a.s. increasing reordering of $(U_1, \ldots, U_{n+1})$ and set $G_n = (e; (U_{1,n}, \ldots, U_{n+1,n}))$. We write $\mathcal{T}_{[n]} = \mathcal{T}(G_n)$ the random real tree spanned by the $n+1$ leaves $\mathbf{p}(U_1), \ldots, \mathbf{p}(U_{n+1})$ and the root and $\tilde{\mathrm{T}}_n = (\mathrm{T}_n; (h_{n,v}, v \in \mathrm{T}_n)) = \tilde{\mathbf{t}}(G_n)$ the associated marked tree. According to Pitman [163], Theorem 7.9 or Aldous [10], the tree $\mathcal{T}_{[n]}$ can also be obtained by stick-breaking procedure or Poisson line-breaking construction. For $1 \leq k \leq n+1$, we denote by $u(U_k)$ the leaf in T_n corresponding to the leaf $\mathbf{p}(U_k)$ in $\mathcal{T}_{[n]}$. See Figure (3.1) for an example with $n = 4$. It is well known that T_n is uniform among the discrete full binary ordered trees with n internal nodes.

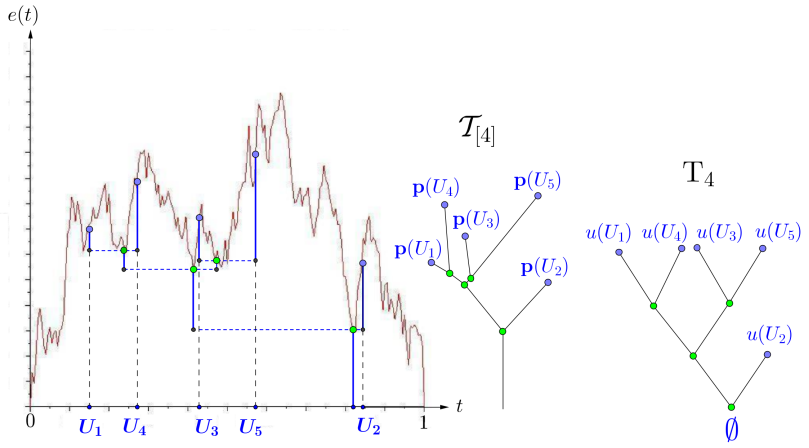


Figure 3.1 – The Brownian excursion, $\mathcal{T}_{[n]}$ and T_n (for $n = 4$).

3.2.5 Simply generated random tree

The presentation of simply generated trees is common in combinatorics. The tools involved in our proofs use Galton-Watson trees. For these reasons, we recall the link between simply generated trees and Galton-Watson trees (see also the survey of Janson [114] for more details). We consider a weight sequence $\mathfrak{p} = (\mathfrak{p}(k), k \in \mathbb{N})$ of non-negative real numbers such that $\sum_{k\in\mathbb{N}} \mathfrak{p}(k) > \mathfrak{p}(1) + \mathfrak{p}(0)$ and $\mathfrak{p}(0) > 0$. For $\mathbf{t} \in \mathbb{T}$, we define its weight as:

$$w(\mathbf{t}) = \prod_{v\in\mathbf{t}} \mathfrak{p}(k_v(\mathbf{t})).$$

We set $w(\mathbb{T}^{(p)}) = \sum_{\mathbf{t}\in\mathbb{T}^{(p)}} w(\mathbf{t})$. For $p \in \mathbb{N}^*$ such that $w(\mathbb{T}^{(p)}) > 0$, a simply generated tree taking values in $\mathbb{T}^{(p)}$ with weight sequence $\mathfrak{p}$ is a $\mathbb{T}^{(p)}$-random variable $\tau^{(p)}$ whose distribution

is characterized by, for all $\mathbf{t} \in \mathbb{T}^{(p)}$:

$$\mathbb{P}(\tau^{(p)} = \mathbf{t}) = \frac{w(\mathbf{t})}{w(\mathbb{T}^{(p)})}.$$

Let $g_{\mathfrak{p}}$ be the generating function of $\mathfrak{p}$: $g_{\mathfrak{p}}(\theta) = \sum_{k\in\mathbb{N}} \theta^k \mathfrak{p}(k)$ for $\theta > 0$. From now on, we assume there exists $\theta > 0$ such that $g_{\mathfrak{p}}(\theta)$ is finite. For $q > 0$ such that $g_{\mathfrak{p}}(q) < +\infty$, let $\mathfrak{p}_q$ be the probability distribution with generating function $\theta \mapsto g_{\mathfrak{p}}(q\theta)/g_{\mathfrak{p}}(q)$. According to [121] see also [3], the distribution of the GW tree τ with offspring distribution $\mathfrak{p}_q$ conditioned on $\{|\tau| = p\}$ is the distribution of $\tau^{(p)}$ and thus does not depend on q. It is easy to check there exists at most one positive root, say $q_{\mathfrak{p}}$, of the equation $g_{\mathfrak{p}}(q) = qg'_{\mathfrak{p}}(q)$. We say that $\mathfrak{p}$ is generic (for the total progeny) if such root $q_{\mathfrak{p}}$ exists and non-generic otherwise. In particular, all weight sequences such that there exists $q > 0$ with $g_{\mathfrak{p}}(q)$ finite and $g_{\mathfrak{p}}(q) < qg'_{\mathfrak{p}}(q)$ (that is $\mathfrak{p}_q$ is a super-critical offspring distribution), are generic.

From now on, we shall assume that $\mathfrak{p}$ is generic. Without loss of generality, by replacing $\mathfrak{p}$ by the probability distribution with generating function $\theta \mapsto g_{\mathfrak{p}}(q_{\mathfrak{p}}\theta)/g_{\mathfrak{p}}(q_{\mathfrak{p}})$, we will assume that $\mathfrak{p}$ is a critical probability distribution, that is:

$$\sum_{k\in\mathbb{N}} \mathfrak{p}(k) = \sum_{k\in\mathbb{N}} k\mathfrak{p}(k) = 1.$$

We recall that $\tau^{(p)}$ is distributed as a critical GW tree τ with offspring distribution $\mathfrak{p}$ conditioned on $\{|\tau| = p\}$, as for all finite tree $\mathbf{t}$, $\mathbb{P}(\tau = \mathbf{t}) = w(\mathbf{t})$.

Local limits for critical GW trees conditioned on having a large total progeny go back to [121] for the generic case (infinite spine case) and [114] for the non-generic case (condensation case), see also [3, 4] and reference therein for more general conditionings. Scaling limits or global limits for GW tree conditioned on having a large total progeny have been studied in [70] for forests (that is collection of GW trees) and in [69, 130] for critical GW tree in the domain of attraction of Lévy trees, see also [129] for more general conditioning of GW trees and [131] for non-generic cases.

3.3 Catalan model

Let $\alpha > 0$ and recall $e = \sqrt{2/\alpha}\, B$, where $B = (B_t, t \in [0,1])$ denotes the normalized Brownian excursion. We also recall that the discrete binary tree T_n, defined in Section 3.2.4 from the Brownian tree $\mathcal{T}_e$, is uniformly distributed among the full ordered rooted binary trees with n internal nodes. In particular, we have $|\mathrm{T}_n| = 2n + 1$. For $n \in \mathbb{N}^*$, we define the weighted random measure A_n on $[0,1]$ defined by $A_n = |\mathrm{T}_n|^{-3/2}\, \mathcal{A}_{\mathrm{T}_n}$, that is for $f \in \mathcal{B}([0,1])$:

$$A_n(f) = |\mathrm{T}_n|^{-3/2} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|\; f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right), \tag{3.17}$$

where $\mathrm{T}_{n,v} = (\mathrm{T}_n)_v$ is the sub-tree of T_n "above" v. Notice that $A_n(\{0\}) = 0$. The next result is proved in Section 3.6.

Theorem 3.4. *We have that a.s. for all $f \in \mathcal{B}([0,1])$, continuous on $(0,1]$ and such that $\lim_{x\to 0+} x^a f(x) = 0$ for some $a \in [0, 1/2)$:*

$$A_n(f) \xrightarrow[n\to+\infty]{} \sqrt{2\alpha}\, \Phi_e(f).$$

We deduce from this Theorem that $(A_n, n \in \mathbb{N}^*)$ converges a.s. for the weak topology towards $\sqrt{2\alpha}\, \Phi_e$.

By convention, for $a \in \mathbb{R}$, we denote the function $x \mapsto x^a \mathbf{1}_{(0,1]}(x)$ defined on $[0,1]$ by x^a. We consider the random variable $Z_\beta = \Phi_e(x^{\beta-1})$, see definition (3.16). The behavior of this random variable and its first moment are given in the following Lemma, whose short proof is given in Remark 3.18.

Lemma 3.5. *We have that a.s. for all $1/2 \geq \beta > 0$, $Z_\beta = +\infty$. We have that a.s. for all $\beta > 1/2$, Z_β is finite and*

$$\mathbb{E}\left[Z_\beta\right] = \frac{1}{2\sqrt{\alpha}} \frac{\Gamma\left(\beta - \frac{1}{2}\right)}{\Gamma(\beta)}. \tag{3.18}$$

We also have the representation formulas $Z_1 = \int_0^1 e(s)\, ds$ and for $\beta > 1$:

$$Z_\beta = \frac{1}{2}\beta(\beta-1)\int_{[0,1]^2} |t-s|^{\beta-2}\, m(s,t)\, ds\, dt. \tag{3.19}$$

We get the following convergence.

Corollary 3.6. *We have that a.s. for all $\beta > 0$,*

$$\lim_{n\to+\infty} |\mathrm{T}_n|^{-(\beta+\frac{1}{2})} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta = \sqrt{2\alpha}\, Z_\beta.$$

Proof. Notice that $|\mathrm{T}_n|^{-(\beta+\frac{1}{2})} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta = A_n(x^{\beta-1})$. For $\beta > 1/2$, the Corollary is then a direct consequence of Theorem 3.4 with $f = x^{\beta-1}$. We now consider the case $1/2 \geq \beta > 0$. Let $c > 0$. Using Theorem 3.4, we have that a.s.:

$$\liminf_{n\to+\infty} A_n(x^{\beta-1}) \geq \lim_{n\to+\infty} A_n(c \wedge x^{\beta-1}) = \sqrt{2\alpha}\, \Phi_e(c \wedge x^{\beta-1}).$$

Letting c goes to infinity, and using that, by Lemma 3.5, $\Phi_e(x^{\beta-1}) = Z_\beta = +\infty$ a.s., we get that a.s. $\liminf_{n\to+\infty} A_n(x^{\beta-1}) \geq \sqrt{2\alpha}\, Z_\beta = +\infty$. Then use a monotonicity argument in β to deduce the results holds a.s. for all $\beta \in (0, 1/2]$. □

Remark 3.7. All the moments of Z_β, for $\beta > 1/2$, are given in [89] (see Proposition 3.5 therein), thanks to the identification provided by Corollary 3.6. The representation formula (3.19) for Z_β is motivated by the formulation of our Corollary 3.6 given in [89] and [88].

Remark 3.8. Corollary 3.6 gives directly that $(|\mathrm{T}_n|^{-3/2} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|, |\mathrm{T}_n|^{-5/2} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}|^2)$ is asymptotically distributed as $\sqrt{2\alpha}\,(Z_1, Z_2)$. Recall $\chi(\mathbf{t})$ defined in (3.6). According to Lemma 3 of [27] or [93], there exists a finite constant K such that, for all $n \geq 3$, we have $\mathbb{E}[\min(|\mathrm{T}_{n,1}|, |\mathrm{T}_{n,2}|)] \leq K|\mathrm{T}_n|^{1/2}$. Since conditionally on $\{v \in \mathrm{T}_n\}$ and $|\mathrm{T}_{n,v}|$, we have that $\mathrm{T}_{n,v}$ is uniformly distributed on the trees with $|\mathrm{T}_{n,v}|$ nodes, we deduce that $\mathbb{E}[\chi(\mathrm{T}_n)] \leq K\mathbb{E}[\mathcal{A}_{\mathrm{T}_n}(\sqrt{x})]$. According to Theorem 3.8 in [89], we have $\mathbb{E}[\mathcal{A}_{\mathrm{T}_n}(\sqrt{x})] = O(n\log(n))$ and thus $\mathbb{E}[\chi(\mathrm{T}_n)] = O(n\log(n))$. Noticing that $\chi(\mathrm{T}_n)$ is non-decreasing in n, using Borel-Cantelli lemma and arguments on convergence determining of measures [4](see proof of Theorem 3.4 in Section 3.6 for a detailed proof in the same spirit), we deduce that a.s. $\lim_{n\to+\infty} |\mathrm{T}_n|^{-3/2}\chi(\mathrm{T}_n) = 0$. Then, we can directly recover the joint asymptotic distribution of the total length path, the Wiener, Sackin, Colless and cophenetic indexes defined by (3.2) in Section 3.1.1 and (3.7) in Section 3.1.2 for the Catalan model. More precisely, we have:

$$\left(\frac{P(\mathrm{T}_n)}{|\mathrm{T}_n|^{3/2}}, \frac{W(\mathrm{T}_n)}{|\mathrm{T}_n|^{5/2}}, \frac{S(\mathrm{T}_n)}{|\mathrm{T}_n|^{3/2}}, \frac{C(\mathrm{T}_n)}{|\mathrm{T}_n|^{3/2}}, \frac{\mathrm{Co}(\mathrm{T}_n)}{|\mathrm{T}_n|^{5/2}}\right) \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{2\alpha}\left(Z_1, 2(Z_1 - Z_2), \frac{Z_1}{2}, \frac{Z_1}{2}, \frac{Z_2}{4}\right).$$

Remark 3.9. We complete Corollary 3.6 by considering the additive functionals $\tilde{F}$, see definition (3.5) used in [89], instead F defined by (3.4). For $\mathbf{t} \in \mathbb{T}$ and $|\mathbf{t}| > 1$, recall $\mathbf{t}^* = \mathbf{t} \setminus \mathcal{L}(\mathbf{t})$ is the tree $\mathbf{t}$ without its leaves. We have that a.s. for all $\beta > 0$,

$$\lim_{n\to+\infty} |\mathrm{T}_n^*|^{-(\beta+\frac{1}{2})} \sum_{v\in\mathrm{T}_n^*} |\mathrm{T}_{n,v}^*|^\beta = 2\sqrt{\alpha}\, Z_\beta. \tag{3.20}$$

[4] i.e. using the characterization of random measures.

This result differs from Corollary 3.6 as $\sqrt{2}$ is replaced by 2. To prove (3.20), first notice that for a full binary tree $|\mathbf{t}^*| = |\mathbf{t}| - |\mathcal{L}(\mathbf{t})| = (|\mathbf{t}| - 1)/2$ so that:

$$|\mathrm{T}_n^*|^{-(\beta+\frac{1}{2})} \sum_{v\in \mathrm{T}_n^*} |\mathrm{T}_{n,v}^*|^\beta = \sqrt{2}\,(|\mathrm{T}_n| - 1)^{-(\beta+\frac{1}{2})} \sum_{v\in \mathrm{T}_n} (|\mathrm{T}_{n,v}| - 1)^\beta.$$

Let $x_+ = \max(x, 0)$ denote the positive part of $x \in \mathbb{R}$. We have $x^\beta \geq (x-1)^\beta \geq x^\beta - c_\beta x^{(\beta-1)_+}$ for all $x \geq 1$ with $c_\beta = 1$ if $0 < \beta \leq 1$ and $c_\beta = \beta$ if $\beta \geq 1$. Then use Corollary 3.6 (two times) to deduce that a.s. for all $\beta > 0$:

$$\lim_{n\to+\infty} |\mathrm{T}_n^*|^{-(\beta+\frac{1}{2})} \sum_{v\in \mathrm{T}_n^*} |\mathrm{T}_{n,v}^*|^\beta = \sqrt{2} \lim_{n\to+\infty} |\mathrm{T}_n|^{-(\beta+\frac{1}{2})} \sum_{v\in \mathrm{T}_n} |\mathrm{T}_{n,v}|^\beta = 2\sqrt{\alpha}\, Z_\beta.$$

The next proposition, whose proof is given in Section 3.7, gives the fluctuations corresponding to the invariance principles of Theorem 3.4. Notice the speed of convergence in the invariance principle is of order $|\mathrm{T}_n|^{-1/4}$.

Proposition 3.10. *Let $f \in \mathcal{C}([0,1])$ be locally Lipschitz continuous on $(0,1]$ with $\| x^a f' \|_{esssup}$ finite for some $a \in (0,1)$. We have the following convergence in distribution:*

$$\left(|\mathrm{T}_n|^{1/4}(A_n - \sqrt{2\alpha}\,\Phi_e)(f),\, A_n\right) \xrightarrow[n\to\infty]{(d)} \left((2\alpha)^{1/4}\sqrt{\Phi_e(xf^2)}\; G,\, \sqrt{2\alpha}\;\Phi_e\right),$$

where G is a standard (centered reduced) Gaussian random variable independent of the excursion e.

Notice the fluctuations for the a.s. convergence towards Z_β with $\beta \geq 1$, given in Corollary 3.6, have an asymptotic variance (up to a multiplicative constant) given by $Z_{2\beta}$.

Remark 3.11. The contribution to the fluctuations is given by the error of approximation of $A_{n,1}(f)$ by $A_{n,2}(f)$, see notations from the proof of Theorem 3.4. This corresponds to the fluctuations coming from the approximation of the branch lengths $(h_{n,v}, v \in \mathrm{T}_n)$ by their mean, which relies on the explicit representation on their joint distribution given in Lemma 3.23. In particular, there is no other contribution to the fluctuations from the approximation of the continuum tree $\mathcal{T}$ by the sub-tree $\mathcal{T}_{[n]}$.

3.4 Simply generated trees model

The main result of this section is Corollary 3.15 in Section 3.4.3. The Sections 3.4.1 and 3.4.2 present the contour process of discrete trees and its convergence towards the contour process of a continuous random tree.

We keep notations from Section 3.2.5 on simply generated random tree. We assume the weight sequence $\mathfrak{p} = (\mathfrak{p}(k), k \in \mathbb{N})$ of non-negative real numbers such that $\sum_{k\in\mathbb{N}} \mathfrak{p}(k) > \mathfrak{p}(1) + \mathfrak{p}(0)$ and $\mathfrak{p}(0) > 0$ is generic. As stated in Section 3.2.5, without loss of generality, we will assume that $\mathfrak{p}$ is a critical probability distribution, that is:

$$\sum_{k\in\mathbb{N}} \mathfrak{p}(k) = \sum_{k\in\mathbb{N}} k\mathfrak{p}(k) = 1.$$

3.4.1 Contour process

Let $\mathbf{t} \in \mathbb{T}$ be a finite tree. The contour process $C^{\mathbf{t}} = (C^{\mathbf{t}}(s), s \in [0, 2|\mathbf{t}|])$ is defined as the distance to the root of a particle visiting continuously each edge of $\mathbf{t}$ at speed one (where all edges are of length 1) according to the lexicographic order of the nodes. More precisely, we set $\emptyset = u(0) < u(1) < \ldots < u(|\mathbf{t}| - 1)$ the nodes of $\mathbf{t}$ ranked in the lexicographic order. By convention, we set $u(|\mathbf{t}|) = \emptyset$.

We set $\ell_0 = 0$, $\ell_{|\mathbf{t}|+1} = 2$ and for $k \in \{1, \dots, |\mathbf{t}|\}$, $\ell_k = d(u(k-1), u(k))$. We set $L_k = \sum_{i=0}^{k} \ell_i$ for $k \in \{0, \dots, |\mathbf{t}|+1\}$, and $L'_k = L_k + d(u(k), \mathfrak{m}(u(k), u(k+1)))$ for $k \in \{0, \dots, |\mathbf{t}|-1\}$. (Notice that $L'_k = L_k$ if and only if $u(k) \prec u(k+1)$.) We have $L_{|\mathbf{t}|} = 2|\mathbf{t}|-2$ and $L_{|\mathbf{t}|+1} = 2|\mathbf{t}|$. We define for $k \in \{0, \dots, |\mathbf{t}| - 1\}$:

- for $s \in [L_k, L'_k)$, the particle goes down from $u(k)$ to $\mathfrak{m}(u(k), u(k+1))$: $C^{\mathbf{t}}(s) = |u(k)| - (s - L_k)$;
- for $s \in [L'_k, L_{k+1})$, the particle goes up from $\mathfrak{m}(u(k), u(k+1))$ to $u(k+1)$: $C^{\mathbf{t}}(s) = |\mathfrak{m}(u(k), u(k+1))| + (s - L'_k)$,

and $C^{\mathbf{t}}(s) = 0$ for $s \in [2|\mathbf{t}| - 2, 2|\mathbf{t}|]$. Notice that $C^{\mathbf{t}}$ is continuous.

For $u \in \mathbf{t}$, we define $\mathcal{I}_u$ the time interval during which the particle explores the edge attached below u. More precisely for $k \in \{1, \dots, |\mathbf{t}| - 1\}$, we set:

$$\mathcal{I}_{u(k)} = [L_k - 1, L_k) \bigcup [L''_k, L''_k + 1),$$

where $L''_k = \inf\{s \geq L_k,\, C^{\mathbf{t}}(s) < |u(k)|\}$ and $\mathcal{I}_\emptyset = [2|\mathbf{t}| - 2, 2|\mathbf{t}|]$. The sets $(\mathcal{I}_u, u \in \mathbf{t})$ are disjoints 2 by 2 with $\bigcup_{u \in \mathbf{t}} \mathcal{I}_u = [0, 2|\mathbf{t}|]$. For $u \in \mathbf{t}$, we have that the Lebesgue measure of $\mathcal{I}_u$ is 2 and

$$C^{\mathbf{t}}(s) \leq d(\emptyset, u) \leq C^{\mathbf{t}}(s) + 1 \quad \text{for all } s \in \mathcal{I}_u. \tag{3.21}$$

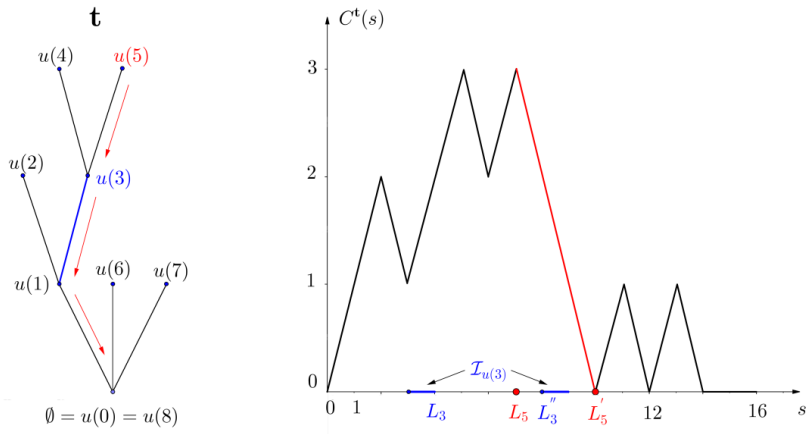


Figure 3.2 – A tree $\mathbf{t}$ with 8 nodes and its contour process $C^{\mathbf{t}}$: for $s \in [L_5, L'_5) = [7, 10)$, the particle goes down from $u(5)$ to $\mathfrak{m}(u(5), u(6)) = \emptyset$; $\mathcal{I}_{u(3)} = [L_3 - 1, L_3) \cup [L''_3, L''_3 + 1) = [4, 5) \cup [8, 9)$ is the time interval during which the particle explores the edge attached below $u(3)$.

3.4.2 Convergence of contour processes

We assume that $\mathfrak{p}$ is a probability distribution on $\mathbb{N}$ such that $1 > \mathfrak{p}(1) + \mathfrak{p}(0) \geq \mathfrak{p}(0) > 0$ and which is critical (that is $\sum_{k \in \mathbb{N}} k\mathfrak{p}(k) = 1$). We also assume that $\mathfrak{p}$ is in the domain of attraction of a stable distribution of Laplace exponent $\psi(\lambda) = \kappa\lambda^\gamma$ with $\gamma \in (1, 2]$ and $\kappa > 0$, and renormalizing sequence $(a_p, p \in \mathbb{N}^*)$ of positive reals: if $(U_k, k \in \mathbb{N}^*)$ are independent random variables with the same distribution $\mathfrak{p}$, and $W_p = \sum_{k=1}^{p} U_k - p$, then W_p/a_p converges in distribution, as p goes to infinity, towards a random variable X with Laplace exponent $-\psi$ (that is $\mathbb{E}[\mathrm{e}^{-\lambda X}] = \mathrm{e}^{\psi(\lambda)}$ for $\lambda \geq 0$). Notice this convergence implies that:

$$\lim_{p \to +\infty} \frac{a_p}{p} = 0. \tag{3.22}$$

Remark 3.12. If $\mathfrak{p}$ has finite variance, say σ^2, then one can take $a_p = \sqrt{p}$ and X is then a centered Gaussian random variable with variance σ^2, so that $\psi(\lambda) = \sigma^2\lambda^2/2$.

The main theorem in Duquesne [69] on the functional convergence in distribution of the contour process stated when $\mathfrak{p}$ is aperiodic, can easily be extended to the case $\mathfrak{p}$ periodic. (Indeed the lack of periodicity hypothesis is mainly used in Lemma 4.5 in [69] which is based on Gnedenko local limit theorem. Since the latter holds *a fortiori* for lattice distributions in the domain of attraction of stable law, it allows to extend the result to such periodic distribution, as soon as one uses sub-sequences on which the conditional probabilities are well defined.) It will be stated in this more general version, see Theorem 3.13 below. Since the contour process is continuous as well as its limit, the convergence in distribution holds on the space $\mathcal{C}([0,1])$ of real continuous functions endowed with the supremum norm.

Theorem 3.13. *Let $\mathfrak{p}$ be a critical probability distribution on $\mathbb{N}$, with $1 > \mathfrak{p}(1) + \mathfrak{p}(0) \geq \mathfrak{p}(0) > 0$, which belongs to the domain of attraction of a stable distribution of Laplace exponent $\psi(\lambda) = \kappa\lambda^\gamma$ with $\gamma \in (1,2]$ and $\kappa > 0$, and renormalizing sequence $(a_p, p \in \mathbb{N}^*)$. Let τ be a GW tree with offspring distribution $\mathfrak{p}$, and $\tau^{(p)}$ be distributed as τ conditionally on $\{|\tau| = p\}$. There exists a random non-negative continuous process $H = (H_s, s \in [0,1])$, such that the following convergence on the space $\mathcal{C}([0,1])$ holds in distribution:*

$$\frac{a_p}{p}\left(C^{\tau^{(p)}}(2ps), s \in [0,1]\right) \xrightarrow[p\to+\infty]{(d)} H,$$

where the convergence is taken along the infinite sub-sequence of p such that $\mathbb{P}(|\tau| = p) > 0$.

The process H, see [69] for a construction of H, is the so called normalized excursion for the height process, introduced in [136], of a Lévy tree with branching mechanism ψ.

Remark 3.14. If $\psi(\lambda) = \alpha\lambda^2$, for some $\alpha > 0$, then H is distributed as $\sqrt{2/\alpha}\, B$, where B is the positive Brownian excursion, see [70].

3.4.3 Main result

The next result is a direct consequence of [69] on the convergence of the contour process of random discrete tree, see Theorem 3.13 given in Section 3.4.2. We keep notations and definitions of Sections 3.4.1 and 3.4.2 below, with H the normalized excursion of the height function associated to the branching mechanism ψ. The proof of the next corollary is given in Section 3.4.4.

Corollary 3.15. *Let $\mathfrak{p}$ be a critical probability distribution on $\mathbb{N}$, with $1 > \mathfrak{p}(1) + \mathfrak{p}(0) \geq \mathfrak{p}(0) > 0$, which belongs to the domain of attraction of a stable distribution of Laplace exponent $\psi(\lambda) = \kappa\lambda^\gamma$ with $\gamma \in (1,2]$ and $\kappa > 0$, and renormalizing sequence $(a_p, p \in \mathbb{N}^*)$. Let τ be a GW tree with offspring distribution $\mathfrak{p}$, and $\tau^{(p)}$ be distributed as τ conditionally on $\{|\tau| = p\}$. We have the following convergence in distribution:*

$$\frac{a_p}{p^2}\mathcal{A}_{\tau^{(p)}} \xrightarrow[p\to+\infty]{(d)} \Phi_H,$$

where we endow the space of non-negative measures with the topology of the weak convergence and where the convergence is taken along the infinite sub-sequence of p such that $\mathbb{P}(|\tau| = p) > 0$.

We set for $\beta > 0$ and $\mathbf{t} \in \mathbb{T}$:

$$Z^*_\beta(\mathbf{t}) = \sum_{v \in \mathbf{t}} |\mathbf{t}_v|^\beta.$$

Using the Skorohod representation theorem, we deduce the following result.

Corollary 3.16. *Assume hypothesis of Corollary 3.15 hold and let Z_β^H be given by (3.15) for $\beta \geq 1$. There exist continuous functions defined on $[1,\infty)$, Θ_p distributed as $\left(\frac{a_p}{p^{\beta+1}} Z_\beta^*(\tau^{(p)}),\ \beta \geq 1\right)$ and Θ distributed as $\left(Z_\beta^H,\ \beta \geq 1\right)$ such that*

$$\Theta_p \xrightarrow[p\to+\infty]{(p.s.)} \Theta$$

for the simple convergence of functions and where the convergence is taken along the infinite sub-sequence of p such that $\mathbb{P}(|\tau| = p) > 0$.

The technical proof of the first part of the next Lemma is given in Section 3.8.6. The second part, which is the representation formula, is a direct consequence of the deterministic Lemma 3.39 in Section 3.8.5 (with $\beta = a + 1$).

Lemma 3.17. *Assume the height function H is associated to the Laplace exponent $\psi(\lambda) = \kappa\lambda^\gamma$ with $\gamma \in (1,2]$ and $\kappa > 0$. We have that a.s. for all $1/\gamma \geq \beta > 0$, $Z_\beta^H = +\infty$, that a.s. for all $\beta > 1/\gamma$, Z_β^H is finite and*

$$\mathbb{E}\left[Z_\beta^H\right] = \frac{1}{\gamma\kappa^{1/\gamma}} \frac{\Gamma\left(\beta - \frac{1}{\gamma}\right)}{\Gamma\left(\beta + 1 - \frac{2}{\gamma}\right)}. \tag{3.23}$$

We also have the representation formulas $Z_1^H = \int_0^1 H(s)\,ds$ and, for $\beta > 1$, $Z_\beta^H = \frac{1}{2}\beta(\beta-1)\int_{[0,1]^2} |t-s|^{\beta-2}\, m_H(s,t)\,ds\,dt$.

Remark 3.18. Lemma 3.5 given in Section 3.3 is a consequence of Lemma 3.17 applied with $H = e$, $\gamma = 2$ and $\kappa = \alpha$.

Remark 3.19. For $\beta \in (0, 1/\gamma]$, we deduce from Corollary 3.15 and Lemma 3.17, using the same arguments as in the proof of Corollary 3.6, the convergence in distribution of the sequence $\left(\frac{a_p}{p^{\beta+1}} Z_\beta^*(\tau^{(p)}),\ p \in \mathbb{N}^* \text{ s.t. } \mathbb{P}(|\tau| = p) > 0\right)$ towards infinity. So the normalization is not relevant to get a proper limit, suggesting we have a "local" regime. The convergence in distribution of this sequence for $\beta \in (1/\gamma, 1)$ towards Z_β^H (which is a.s. finite) is an open question, but we conjecture it holds. This conjecture and Corollary 3.16 would then give that for simply generated trees, under the hypothesis of Corollary 3.15, there is a phase transition at $\beta = 1/\gamma$ between a "global" regime ($\beta > 1/\gamma$) and a "local" regime ($\beta \leq 1/\gamma$).

Remark 3.20. If $\mathfrak{p}$ has finite variance, say σ^2, then one can take $a_p = \sqrt{p}$ in Corollaries 3.15 and 3.16 and H is equal to $(2/\sigma)B$ which corresponds to $\psi(\lambda) = \sigma^2\lambda^2/2$, see Remarks 3.12 and 3.14. By scaling, or using that the limit in Theorem 3.4 does not depend on α, we deduce that in this case $\Phi_H = \frac{2}{\sigma}\Phi_B$ and $Z_\beta^H = \frac{2}{\sigma} Z_\beta^B$ in Corollaries 3.15 and 3.16.

3.4.4 Proof of Corollary 3.15

Elementary functionals of finite trees

Let $\mathbf{t} \in \mathbb{T}$ be a finite tree and $k \in \mathbb{N}^*$. For $\mathbf{u} = (u_1, \ldots, u_k) \in \mathbf{t}^k$, we define $\mathfrak{m}(\mathbf{u}) = \mathfrak{m}(\{u_1, \ldots, u_k\})$ the most recent common ancestor of $u_1, \ldots, u_k$. We consider the following elementary functional of a tree, defined for $\mathbf{t} \in \mathbb{T}$:

$$D_k(\mathbf{t}) = \sum_{\mathbf{u} \in \mathbf{t}^k} d(\emptyset, \mathfrak{m}(\mathbf{u})). \tag{3.24}$$

We have:

$$\sum_{v\in\mathbf{t}} |\mathbf{t}_v|^k = D_k(\mathbf{t}) + |\mathbf{t}|^k, \tag{3.25}$$

which we obtain from the following equalities

$$\sum_{v\in\mathbf{t}} |\mathbf{t}_v|^k = \sum_{v\in\mathbf{t}} \sum_{\mathbf{u}\in\mathbf{t}^k} \mathbf{1}_{\{v\preccurlyeq\mathfrak{m}(\mathbf{u})\}} = \sum_{\mathbf{u}\in\mathbf{t}^k} \sum_{v\in\mathbf{t}} \mathbf{1}_{\{v\preccurlyeq\mathfrak{m}(\mathbf{u})\}} = \sum_{\mathbf{u}\in\mathbf{t}^k} (d(\emptyset,\mathfrak{m}(\mathbf{u}))+1).$$

For $x=(x_1,\dots,x_k)\in\mathbb{R}^k$, denote by $(x_{(1)},\dots,x_{(k)})$ its order statistics which are uniquely defined by $x_{(1)}\le\cdots\le x_{(n)}$ and $\sum_{i=1}^k \delta_{x_i} = \sum_{i=1}^k \delta_{x_{(i)}}$, with δ_z the Dirac mass at z. Recall the notation $m_h(s,t)$, see (3.13), for the minimum of h over the interval with bounds s and t. We set:

$$\mathcal{D}_k(\mathbf{t}) = \int_{[0,|\mathbf{t}|]^k} m_{C^{\mathbf{t}}}(2x_{(1)}, 2x_{(k)})\,dx, \tag{3.26}$$

with the conventions that if $k=1$, then $\mathcal{D}_1(\mathbf{t}) = \int_{[0,|\mathbf{t}|]} C^{\mathbf{t}}(2x)\,dx$.

We have the following lemma.

Lemma 3.21. *We have for $\mathbf{t}\in\mathbb{T}$ and $k\in\mathbb{N}^*$:*

$$0 \le D_k(\mathbf{t}) - \mathcal{D}_k(\mathbf{t}) \le |\mathbf{t}|^k. \tag{3.27}$$

Proof. For $\mathbf{u}=(u_1,\dots,u_k)\in\mathbf{t}^k$, we have the following generalization of (3.21): for $x=(x_1,\dots,x_k)\in\prod_{i=1}^k \mathcal{I}_{u_i}$,

$$m_{C^{\mathbf{t}}}(x_{(1)},x_{(k)}) \le d(\emptyset,\mathfrak{m}(\mathbf{u})) \le m_{C^{\mathbf{t}}}(x_{(1)},x_{(k)}) + 1.$$

(Notice that $m_{C^{\mathbf{t}}}(x_{(1)},x_{(k)}) = d(\emptyset,\mathfrak{m}(\mathbf{u}))$ as soon as $\mathfrak{m}(\mathbf{u}) \prec u_i$ for all $i\in\{1,\dots,k\}$.) We deduce that:

$$0 \le 2^k d(\emptyset,\mathfrak{m}(\mathbf{u})) - \int_{\prod_{i=1}^k \mathcal{I}_{u_i}} m_{C^{\mathbf{t}}}(x_{(1)},x_{(k)})\,dx \le 2^k.$$

By summing over $\mathbf{u}\in\mathbf{t}^k$, we get:

$$0 \le 2^k D_k(\mathbf{t}) - \int_{[0,2|\mathbf{t}|]^k} m_{C^{\mathbf{t}}}(x_{(1)},x_{(k)})\,dx \le 2^k|\mathbf{t}|^k.$$

Use the change of variable $2y=x$ to get (3.27). □

Convergence of additive functionals

We now give the main result of this Section.

Corollary 3.22. *Under the hypothesis and notations of Theorem 3.13, we have the following convergences in distribution for all $k\in\mathbb{N}^*$:*

$$\lim_{p\to+\infty} \frac{a_p}{p^{k+1}} \sum_{v\in\tau^{(p)}} |\tau_v^{(p)}|^k \overset{(d)}{=} \lim_{p\to+\infty} \frac{a_p}{p^{k+1}} \sum_{\mathbf{u}\in(\tau^{(p)})^k} d(\emptyset,\mathfrak{m}(\mathbf{u})) \overset{(d)}{=} \int_0^1 ds \int_0^{H(s)} dr\,\sigma_{r,s}(H)^{k-1},$$

where $\sigma_{r,s}(H)$ is the length of the excursion of the height process H above r straddling s defined in (3.14) and where the convergence is taken along the infinite sub-sequence of p such that $\mathbb{P}(|\tau|=p)>0$.

Proof. Recall notation $m_h(s,t)$ and $\sigma_{r,s}(h)$ given in (3.13) and (3.14). We shall take limits along the infinite sub-sequence of p such that $\mathbb{P}(|\tau| = p) > 0$.

Recall definitions (3.24) of $D_k(\mathbf{t})$ and (3.26) of $\mathcal{D}_k(\mathbf{t})$. Thanks to Lemma 3.21 and (3.22) which implies that $(p^{-(k+1)} a_p(D_k(\tau^{(p)}) - \mathcal{D}_k(\tau^{(p)})),\ p \in \mathbb{N}^*)$ converges in probability towards 0 and to (3.25), we see the proof of the corollary is complete as soon as we obtain that for all $k \in \mathbb{N}^*$:

$$\lim_{p\to+\infty} \frac{a_p}{p^{k+1}} \mathcal{D}_k(\tau^{(p)}) \overset{(d)}{=} \int_0^1 ds \int_0^{H(s)} dr\, \sigma_{r,s}(H)^{k-1}. \tag{3.28}$$

We deduce from Theorem 3.13 the following convergence in law:

$$\frac{a_p}{p^2} \mathcal{D}_1(\tau^{(p)}) = \frac{a_p}{p^2} \int_{[0,p]} C^{\tau^{(p)}}(2x)\, dx = \int_{[0,1]} \frac{a_p}{p} C^{\tau^{(p)}}(2ps)\, ds \xrightarrow[p\to+\infty]{(d)} \int_{[0,1]} ds\, H(s).$$

This gives (3.28) for $k = 1$. We have that for $k \geq 2$ and $\mathbf{t} \in \mathbb{T}$:

$$\begin{aligned} 2\mathcal{D}_k(\mathbf{t}) &= 2 \int_{[0,|\mathbf{t}|]^k} m_{C^{\mathbf{t}}}(2x_{(1)}, 2x_{(k)})\, dx \\ &= k(k-1) \int_{[0,|\mathbf{t}|]^2} dx_1 dx_2\, |x_2 - x_1|^{k-2}\, m_{C^{\mathbf{t}}(2\bullet)}(x_1, x_2) \\ &= k(k-1)|\mathbf{t}|^k \int_{[0,1]^2} dx_1 dx_2\, |x_2 - x_1|^{k-2}\, m_{C^{\mathbf{t}}(2|\mathbf{t}|\bullet)}(x_1, x_2), \end{aligned}$$

where we used (3.26) in the first equality, we choose $x_{(1)}$ and $x_{(k)}$ among $x_1, \ldots, x_k$ for the second one and we used the change of variable x_i to $|\mathbf{t}|x_i$ for the last one. We deduce from Theorem 3.13 the following convergence in law for all $k \in \mathbb{N}^*$ such that $k \geq 2$:

$$\frac{a_p}{p^{k+1}} \mathcal{D}_k(\tau^{(p)}) \xrightarrow[p\to+\infty]{(d)} \frac{k(k-1)}{2} \int_{[0,1]^2} ds ds'\, |s' - s|^{k-2}\, m_H(s, s').$$

Then use (3.57) from Lemma 3.39 to get (3.28). This ends the proof. □

Conclusion

We deduce from the proof of Corollary 3.22, using the Skorohod representation theorem, that all the convergences in distribution of Corollary 3.22 hold simultaneously for all $k \in \mathbb{N}^*$. We thus get that $\lim_{n\to+\infty} \frac{a_p}{p^2} \mathcal{A}_{\tau^{(p)}}(x^{k-1}) \overset{(d)}{=} \Phi_H(x^{k-1})$, simultaneously for all $k \in \mathbb{N}^*$. Since on $[0,1]$, the convergence of moments implies the weak convergence of finite measures, we deduce that the random measure $\frac{a_p}{p^2} \mathcal{A}_{\tau^{(p)}}$ converges in distribution towards Φ_H for the topology of weak convergence of finite measures on $[0,1]$.

3.5 Preliminary Lemmas

Recall $\mathcal{T}$ is the real tree coded by the excursion e, see Section 3.2.3 and $\mathcal{T}_{[n]}$ is the (smallest) sub-tree of $\mathcal{T}_e$ containing $n+1$ leaves picked uniformly at random and the root, see Section 3.2.4. Recall $(\mathrm{T}_n, (h_{n,v}, v \in \mathrm{T}_n))$ denote the corresponding marked tree. Intuitively, for $v \in \mathrm{T}_n$, $h_{n,v}$ is the length of the branch below the branching point with label v in $\mathcal{T}_{[n]}$ (when keeping the order on the leaves). We recall, see [10], [163] (Theorem 7.9) or [70], that the density of $(h_{n,v}, v \in \mathrm{T}_n)$ is, conditionally on T_n, given by:

$$f_n((h_{n,v}, v \in \mathrm{T}_n)) = 2\frac{(2n)!}{n!}\, \alpha^{n+1}\, L_n\, \mathrm{e}^{-\alpha L_n^2} \prod_{v\in\mathrm{T}_n} \mathbf{1}_{\{h_{n,v}>0\}}, \tag{3.29}$$

where $L_n = \sum_{v\in \mathrm{T}_n} h_{n,v}$ denotes the total length of $\mathcal{T}_{[n]}$. Notice that the edge-lengths have an exchangeable distribution and are independent of the shape tree T_n. Furthermore, elementary computations give that $(h_{n,v}, v \in \mathrm{T}_n)$, with $v \in \mathrm{T}_n$ ranked in the lexicographic order, has, conditionally on T_n and L_n, the same distribution as $(L_n\Delta_1, \ldots, L_n\Delta_{2n+1})$, where $\Delta_1, \ldots, \Delta_{2n+1}$ represents the lengths of the $2n+1$ intervals obtained by cutting $[0,1]$ at $2n$ independent uniform random variables on $[0,1]$ and independent of L_n. We thus deduce the following elementary Lemma.

Lemma 3.23. *Conditionally on $\mathrm{T}_n = \mathbf{t}$, the random vector $(h_{n,v}, v \in \mathbf{t})$ has the same distribution as $(L_n E_v/S_{\mathbf{t}}, v \in \mathbf{t})$, where $(E_u, u \in \mathcal{U})$ are independent exponential random variables with mean 1, independent of T_n and L_n, and $S_{\mathbf{t}} = \sum_{v\in\mathbf{t}} E_v$.*

According to [2], we have that a.s. $\lim_{n\to+\infty} L_n/\sqrt{n} = 1/\sqrt{\alpha}$. We then deduce from Lemma 3.23 that $(2n+1)\sqrt{\alpha}\, h_{n,\emptyset}/\sqrt{n}$ converges in distribution towards $E_\emptyset$ as n goes to infinity. Intuitively, we get that $2\sqrt{\alpha n}\,\mathbb{E}[h_{n,v}]$ is of order 1, for $v \in \mathrm{T}_n$. Recall the random measure A_n is defined in (3.17). We introduce the random measure:

$$A_{1,n} = 2\sqrt{\alpha n}\,\mathbb{E}[h_{n,\emptyset}]A_n.$$

Lemma 3.24. *Let $a \in [0,1/2)$. There exists a finite constant C such that for all $f \in \mathcal{B}([0,1])$ and $n \in \mathbb{N}^*$, we have:*

$$\mathbb{E}\left[|A_n(f) - A_{1,n}(f)|\right] \le C\|x^a f\|_\infty\, n^{-1}.$$

Proof. Let $a \in [0,1/2)$ and $f \in \mathcal{B}([0,1])$. Using (3.55) in the Appendix, we deduce that for all $n \in \mathbb{N}^*$, we have $\left|1 - 2\sqrt{\alpha n}\,\mathbb{E}[h_{n,\emptyset}]\right| \le 1/2n$. Using (3.44) in Lemma 3.34, we deduce that:

$$\mathbb{E}[|A_n(f) - A_{1,n}(f)|] \le \frac{1}{2n}\mathbb{E}[|A_n(f)|] \le \frac{C_{1,1-a}}{2n}\,\|x^a f\|_\infty.$$

□

Intuitively, $h_{n,v}$ is of the same order of its expectation. Since the random variables $(h_{n,v}, v \in \mathrm{T}_n)$ are exchangeable, we deduce that $h_{n,v}$ is of the same order as $\mathbb{E}[h_{n,\emptyset}]$. Based on this intuition, we define the random measure $A_{2,n}$ as follows. For $f \in \mathcal{B}([0,1])$, we set:

$$A_{2,n}(f) = 2\sqrt{\alpha n}\;|\mathrm{T}_n|^{-3/2} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) h_{n,v}.$$

Lemma 3.25. *Let $a \in [0,1/2)$. There exists a finite constant C such that for all $f \in \mathcal{B}([0,1])$ and $n \in \mathbb{N}^*$, we have:*

$$\mathbb{E}[|A_{1,n}(f) - A_{2,n}(f)|] \le C\|x^a f\|_\infty\, n^{-1/4}.$$

Proof. Let $a \in [0,1/2)$ and $f \in \mathcal{B}([0,1])$. For $v \in \mathrm{T}_n$, we set $Y_{n,v} = \sqrt{n}(\mathbb{E}[h_{n,v}] - h_{n,v})$ and

$$K_n = \frac{1}{2\sqrt{\alpha}}(A_{1,n}(f) - A_{2,n}(f)) = |\mathrm{T}_n|^{-3/2} \sum_{v\in\mathrm{T}_n} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) Y_{n,v}.$$

Using that $(h_{n,v}, v \in \mathrm{T}_n)$ is exchangeable, elementary computations give:

$$\mathbb{E}\left[K_n^2|\mathrm{T}_n\right] \le |\mathrm{T}_n|^{-1/2} A_n(xf^2)\mathbb{E}[Y_{n,\emptyset}^2] + A_n(|f|)^2|\mathbb{E}[Y_{n,\emptyset}Y_{n,1}]|.$$

Then using (3.44) and (3.45) in Lemma 3.34 and (3.56) in Lemma 3.38, we get:

$$\mathbb{E}[K_n^2] = \mathbb{E}\left[\mathbb{E}[K_n^2|\mathrm{T}_n]\right] \le \frac{C_{1,1}}{2\alpha\sqrt{2n+1}}\|x^{1/2}f\|_\infty^2 + \frac{C_{2,1-a}^2}{8\alpha n}\|x^a f\|_\infty^2 \le \frac{c}{\sqrt{n}}\|x^a f\|_\infty^2,$$

for some finite constant c which does not depend on n and f. □

Let $\mathcal{L}_{n,v} = \{u \in \mathrm{T}_n;\, v \preccurlyeq u,\, k_u(\mathrm{T}_n) = 0\}$ be the set of leaves of T_n with ancestor v, and $|\mathcal{L}_{n,v}|$ be its cardinality. Notice the number of leaves of $\mathrm{T}_{n,v}$ is exactly $|\mathcal{L}_{n,v}|$. We now approximate the multiplying factor $|\mathrm{T}_{n,v}|$ in $A_{2,n}$ by twice the number of leaves in $\mathrm{T}_{n,v}$ as $2|\mathcal{L}_{n,v}| = |\mathrm{T}_{n,v}| + 1$. For this reason, we set for $f \in \mathcal{B}([0,1])$:

$$A_{3,n}(f) = 4\sqrt{\alpha n}\ |\mathrm{T}_n|^{-3/2} \sum_{v \in \mathrm{T}_n} |\mathcal{L}_{n,v}|\, f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) h_{n,v}.$$

Lemma 3.26. *Let $a \in [0, 1/2)$. For all $f \in \mathcal{B}([0,1])$ and $n \in \mathbb{N}^*$, we have:*

$$\mathbb{E}[|A_{2,n}(f) - A_{3,n}(f)|] \leq \|x^a f\|_\infty\, n^{a-\frac{1}{2}}.$$

Proof. Let $a \in [0, 1/2)$ and $f \in \mathcal{B}([0,1])$. Since $2|\mathcal{L}_{n,v}| = |\mathrm{T}_{n,v}| + 1$, we get that:

$$|A_{2,n}(f) - A_{3,n}(f)| \leq 2\sqrt{\alpha n}\, |\mathrm{T}_n|^{-3/2} \sum_{v \in \mathrm{T}_n} |f|\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) h_{n,v}.$$

As $|\mathrm{T}_{n,v}| \geq 1$ and $a \geq 0$, we get that $|f|\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) \leq \|x^a f\|_\infty\, |\mathrm{T}_n|^a$. We deduce that:

$$|A_{2,n}(f) - A_{3,n}(f)| \leq 2\sqrt{\alpha n}\, L_n |\mathrm{T}_n|^{a-\frac{3}{2}}\, \|x^a f\|_\infty.$$

According to (3.54), we have $2\sqrt{\alpha n}\,\mathbb{E}[L_n] \leq |\mathrm{T}_n|$. We deduce that $\mathbb{E}[|A_{2,n}(f) - A_{3,n}(f)|] \leq |\mathrm{T}_n|^{a-\frac{1}{2}}\, \|x^a f\|_\infty$. □

We define $\mathcal{N}_{n,r,U_k}$ as the number of leaves of the sub-tree $\mathcal{T}_{[n]}$ which are distinct from $\mathbf{p}(U_k)$ and such that their most recent common ancestor with $\mathbf{p}(U_k)$ is at distance further than r from the root. More precisely, using the definition (3.13) of m, we have:

$$\mathcal{N}_{n,r,U_k} + 1 = \mathrm{Card}\,\{i \in \{1, \ldots, n+1\},\, m(U_i, U_k) \geq r\}.$$

In particular, we deduce from the construction of $\mathcal{T}_{[n]}$ and T_n that for $1 \leq k \leq n+1$:

$$\sum_{v \preccurlyeq u(U_k)} f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) h_{n,v} = \int_0^{e(U_k)} dr\, f\left(\frac{2\mathcal{N}_{n,r,U_k} + 1}{2n+1}\right), \tag{3.30}$$

where $u(U_k)$ is the leaf in T_n corresponding to the leaf $\mathbf{p}(U_k)$ in $\mathcal{T}_{[n]}$.

Recall that, for $v \in \mathrm{T}_n$, $\mathcal{L}_{n,v}$ denotes the set of leaves of T_n with ancestor v and $\mathcal{L}(\mathrm{T}_n) = \mathcal{L}_{n,\emptyset}$ denotes the set of leaves of T_n. We deduce that:

$$\begin{aligned} A_{3,n}(f) &= 4\sqrt{\alpha n}\ |\mathrm{T}_n|^{-3/2} \sum_{v \in \mathrm{T}_n} |\mathcal{L}_{n,v}|\, f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) h_{n,v} \\ &= 4\sqrt{\alpha n}\ |\mathrm{T}_n|^{-3/2} \sum_{u \in \mathcal{L}(\mathrm{T}_n)} \sum_{v \preccurlyeq u} f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) h_{n,v} \\ &= 4\sqrt{\alpha n}\ |\mathrm{T}_n|^{-3/2} \sum_{k=1}^{n+1} \int_0^{e(U_k)} dr\, f\left(\frac{2\mathcal{N}_{n,r,U_k} + 1}{2n+1}\right), \end{aligned}$$

where we used (3.30) for the last equality. Notice that by construction, conditionally on e and U_k, the random variable $\mathcal{N}_{n,r,U_k}$ is binomial with parameter (n, σ_{r,U_k}). For this reason, we consider the following approximation of $A_{3,n}(f)$. For $f \in \mathcal{B}([0,1])$ non-negative, we set:

$$A_{4,n}(f) = 4\sqrt{\alpha n}\ |\mathrm{T}_n|^{-3/2} \sum_{k=1}^{n+1} \int_0^{e(U_k)} dr\, f(\sigma_{r,U_k}).$$

Lemma 3.27. *We have the following properties.*

(i) For $a \in (0,1)$, there exists a finite constant $C(a)$ such that if $f \in \mathcal{B}([0,1])$ is locally Lipschitz continuous on $(0,1]$, we have for all $n \in \mathbb{N}^*$:

$$\mathbb{E}[|A_{3,n}(f) - A_{4,n}(f)|] \leq C(a) \, \| x^a f' \|_{esssup} \, n^{-1/2}.$$

(ii) If $a \in (-1/2, 0]$, there exists a finite constant $C(a)$ such that we have for all $n \in \mathbb{N}^*$:

$$\mathbb{E}[|A_{3,n}(x^a) - A_{4,n}(x^a)|] \leq C(a) \, n^{-(2a+1)/8}.$$

Remark 3.28. We can extend (i) of Lemma 3.27 to get that for a uniformly Hölder continuous function f with exponent $\lambda > 1/2$, we have $\mathbb{E}[|A_{3,n}(f) - A_{4,n}(f)|] = O(n^{-\lambda/2})$. This allows the extension of Proposition 3.10 to such functions.

Proof. For $s \in [0,1]$, let $\mathcal{N}_{n,r,s}$ be a random variable which is, conditionally on e, binomial with parameter $(n, \sigma_{r,s})$. Notice, this is consistent with the definition of $\mathcal{N}_{n,r,U_k}$. Hence we get, for $f \in \mathcal{B}([0,1])$,

$$\begin{aligned}\mathbb{E}\left[|A_{3,n}(f) - A_{4,n}(f)|\right] &\leq 4\sqrt{\alpha n}|\mathrm{T}_n|^{-\frac{3}{2}} \sum_{k=1}^{n+1} \mathbb{E}\left[\int_0^{e(U_k)} \left| f\left(\frac{2\mathcal{N}_{n,r,U_k}+1}{2n+1}\right) - f(\sigma_{r,U_k})\right| dr\right] \\ &\leq 4\sqrt{\alpha}\int_0^1 ds\, \mathbb{E}\left[\int_0^{e(s)} dr\, \mathbb{E}\left[\left| f\left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right) - f(\sigma_{r,s})\right| \mid e\right] dr\right]. \end{aligned} \tag{3.31}$$

We first prove property (i). Let $a \in (0,1)$ and $f \in \mathcal{B}([0,1])$ be locally Lipschitz continuous on $(0,1]$. Using (ii) of Lemma 3.35, we have that for $s \in (0,1)$ and $r \in (0, e(s))$,

$$\mathbb{E}\left[\left| f\left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right) - f(\sigma_{r,s})\right| \mid e\right] \leq \frac{\| x^a f' \|_{\mathrm{esssup}}}{1-a}\left(\sigma_{r,s}^{-\frac{a}{2}} + \sigma_{r,s}^{\frac{1}{2}-a}\right) n^{-1/2}. \tag{3.32}$$

We recall that $Z_\beta = \int_0^1 ds \int_0^{e(s)} dr\, \sigma_{r,s}^{\beta-1}$ for $\beta > 0$. Thus, we have $\mathbb{E}\left[Z_{\frac{3}{2}-a}\right] \leq \mathbb{E}\left[Z_{1-\frac{a}{2}}\right]$; the last term being finite thanks to Lemma 3.5. We deduce from (3.31) and (3.32) that

$$\mathbb{E}\left[|A_{3,n}(f) - A_{4,n}(f)|\right] \leq 8\sqrt{\alpha}\frac{\| x^a f' \|_{\mathrm{esssup}}}{1-a}\, \mathbb{E}\left[Z_{1-\frac{a}{2}}\right] n^{-1/2}.$$

This achieves the proof of property (i).

We now prove property (ii). We consider $a \in (-1/2, 0)$ and $f(x) = x^a$, as the case $a = 0$ is obvious. Let $\gamma > 0$. We write:

$$\int_0^1 ds\, \mathbb{E}\left[\int_0^{e(s)} dr\, \mathbb{E}\left[\left|\left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right)^a - \sigma_{r,s}^a\right| \mid e\right]\right] = \kappa_1 + \kappa_2 + \kappa_3,$$

with $\kappa_i = \int_0^1 ds\, \mathbb{E}\left[\int_0^{e(s)} dr\, \mathbb{E}\left[\mathbf{1}_{D_i}\left|\left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right)^a - \sigma_{r,s}^a\right| \mid e\right]\right]$ and:

$$D_1 = \left\{\sigma_{r,s} > 2n^{-\gamma}, \frac{2\mathcal{N}_{n,r,s}+1}{2n+1} > n^{-\gamma}\right\}, \quad D_2 = \left\{\sigma_{r,s} > 2n^{-\gamma}, \frac{2\mathcal{N}_{n,r,s}+1}{2n+1} \leq n^{-\gamma}\right\},$$

and $D_3 = (D_1 \bigcup D_2)^c$. We start considering κ_1. Notice that, thanks to (3.49) with $b = 1+a$, we have $|x^a - y^a| \leq x^a y^{-1}|x-y| \leq n^{\gamma(1-a)}|x-y|$ if $x, y \in [n^{-\gamma}, +\infty)$. Using this inequality with $x = \frac{2\mathcal{N}_{n,r,s}+1}{2n+1}$ and $y = \sigma_{r,s}$, we obtain:

$$\kappa_1 \leq n^{\gamma(1-a)} \int_0^1 ds\, \mathbb{E}\left[\int_0^{e(s)} \mathbb{E}\left[\left|\frac{2\mathcal{N}_{n,r,s}+1}{2n+1} - \sigma_{r,s}\right| \,|\, e\right] dr\right]. \tag{3.33}$$

Moreover, if X is a binomial random variable with parameter (n,p), then we have:

$$\mathbb{E}\left[\left|\frac{2X+1}{2n+1} - p\right|\right]^2 \leq \mathbb{E}\left[\left(\frac{2X+1}{2n+1} - p\right)^2\right] \leq \frac{1}{2n+1} \leq \frac{1}{n}.$$

With $X = \mathcal{N}_{n,r,s}$ and $p = \sigma_{r,s}$, we get that $\mathbb{E}\left[\left|\frac{2\mathcal{N}_{n,r,s}+1}{2n+1} - \sigma_{r,s}\right| \,|\, e\right] \leq \frac{1}{\sqrt{n}}$. We deduce from (3.33) that:

$$\kappa_1 \leq \mathbb{E}\left[\int_0^1 e(s)\, ds\right] n^{\gamma(1-a)-\frac{1}{2}}. \tag{3.34}$$

We give an upper bound of κ_2. We first recall Hoeffding's inequality: if X is a binomial random variable with parameter (n,p), and $t > 0$, then we have $\mathbb{P}(np - X > nt) \leq \exp(-2nt^2)$. Using that $\{p - \frac{2X+1}{2n+1} > n^{-\gamma}\} \subset \{np - X > n^{1-\gamma}\}$, we deduce that:

$$\mathbb{P}\left(p - \frac{2X+1}{2n+1} > n^{-\gamma}\right) \leq \mathbb{P}\left(np - X > n^{1-\gamma}\right) \leq \exp\left(-2n^{1-2\gamma}\right). \tag{3.35}$$

Notice that on D_2, we have $0 \leq \left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right)^a - \sigma_{r,s}^a \leq \left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right)^a \leq (2n+1)^{-\gamma a}$ as well as $\sigma_{r,s} - \frac{2\mathcal{N}_{n,r,s}+1}{2n+1} > n^{-\gamma}$. Hence, we obtain:

$$\begin{aligned}\kappa_2 &\leq (2n+1)^{-\gamma a} \int_0^1 ds\, \mathbb{E}\left[\int_0^{e(s)} \mathbb{P}\left(\sigma_{r,s} - \frac{2\mathcal{N}_{n,r,s}+1}{2n+1} > n^{-\gamma} \middle| e\right) dr\right] \\ &\leq \mathbb{E}[Z_1]\,(2n+1)^{-\gamma a}\, \mathrm{e}^{-2n^{1-2\gamma}}.\end{aligned} \tag{3.36}$$

Finally, we consider κ_3. Let $\eta \in (0, a+1/2)$. We have:

$$\begin{aligned}&\mathbb{E}\left[\int_0^{e(s)} \mathbf{1}_{\{\sigma_{r,s} \leq 2n^{-\gamma}\}} \mathbb{E}\left[\left|\left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right)^a - \sigma_{r,s}^a\right| \,|\, e\right] dr\right] \\ &\qquad \leq \mathbb{E}\left[\int_0^{e(s)} \mathbf{1}_{\{\sigma_{r,s} \leq 2n^{-\gamma}\}} \mathbb{E}\left[\left(\frac{2\mathcal{N}_{n,r,s}+1}{2n+1}\right)^a + \sigma_{r,s}^a \,|\, e\right] dr\right] \\ &\qquad \leq 3\mathbb{E}\left[\int_0^{e(s)} \mathbf{1}_{\{\sigma_{r,s} \leq 2n^{-\gamma}\}} \sigma_{r,s}^a\, dr\right] \\ &\qquad \leq 3 \cdot 2^\eta n^{-\gamma\eta}\, \mathbb{E}\left[\int_0^{e(s)} dr\, \sigma_{r,s}^{a-\eta}\right],\end{aligned}$$

where we used (i) of Lemma 3.35 for the second inequality. Recall that $D_3 = \{\sigma_{r,s} \leq 2n^{-\gamma}\}$. We deduce that:

$$\kappa_3 \leq \int_0^1 ds\, 3 \cdot 2^\eta n^{-\gamma\eta}\, \mathbb{E}\left[\int_0^{e(s)} dr\, \sigma_{r,s}^{a-\eta}\right] = 3 \cdot 2^\eta n^{-\gamma\eta}\, \mathbb{E}[Z_{a-\eta+1}]. \tag{3.37}$$

Choose $\gamma = 1/3$ and $\eta = 3(2a+1)/8$. Thanks to Lemma 3.5, we get that $\mathbb{E}\left[\int_0^1 e(s)\, ds\right] = \mathbb{E}[Z_1]$ is finite and that $\mathbb{E}[Z_{a-\eta+1}]$ is also finite since $a - \eta + 1 > 1/2$ as $a > -1/2$. Therefore,

we deduce from (3.31) and then (3.34), (3.36) and (3.37) that there exists a finite constant $C(a)$ such that we have for all $n \in \mathbb{N}^*$:

$$\mathbb{E}[|A_{3,n}(x^a) - A_{4,n}(x^a)|] \leq C(a)\, n^{-(2a+1)/8}.$$

□

Lemma 3.29. *For all $f \in \mathcal{B}([0,1])$ such that $f \geq 0$ and $\|x^a f\|_\infty < +\infty$ for some $a \in [0, 1/2)$, we have:*

$$A_{4,n}(f) \xrightarrow[n\to+\infty]{a.s.} \sqrt{2\alpha}\ \Phi_e(f).$$

Proof. Let $f \in \mathcal{B}([0,1])$ such that $f \geq 0$ and $\|x^a f\|_\infty < +\infty$ for some $a \in [0, 1/2)$. Let U be uniform on $[0,1]$ and independent of e. Recall $Z_\beta = \int_0^1 ds \int_0^1 dr\, \sigma_{r,s}^{\beta-1}$ defined in (3.15). Notice that:

$$\mathbb{E}\Big[\int_0^{e(U)} dr\, f(\sigma_{r,U}) \,\Big|\, e \Big] \leq \| x^a f \|_\infty\, Z_{1-a}.$$

Since $1 - a > 1/2$, we deduce from Lemma 3.5 that a.s. $Z_{1-a} < +\infty$. Then, use the strong law of large numbers (conditionally on e) to deduce that $A_{4,n}(f)$ converges a.s. towards $\sqrt{2\alpha}\ \Phi_e(f)$ as n goes to infinity. □

3.6 Proof of Theorem 3.4

Let $a > -1/2$. According to Lemmas 3.24, 3.25, 3.26 and 3.27 (use (i) for $a > 0$ and (ii) for $a \in (-1/2, 0]$), there exists $\varepsilon > 0$ and a finite constant c such that for all $n \in \mathbb{N}^*$, we have $\mathbb{E}[|A_n(x^a) - A_{4,n}(x^a)|] \leq c n^{-\varepsilon}$. Since according to Lemma 3.29, we have a.s. that $\lim_{n\to+\infty} A_{4,n}(x^a) = \sqrt{2\alpha}\ \Phi_e(x^a)$, we deduce from Borel-Cantelli lemma that, with $\varphi(n) = \lceil n^{2/\varepsilon} \rceil$, we have a.s. $\lim_{n\to+\infty} A_{\varphi(n)}(x^a) = \sqrt{2\alpha}\ \Phi_e(x^a)$.

For $n' \geq n \geq 1$, we have $\mathcal{T}_{[n]} \subset \mathcal{T}_{[n']}$. Unfortunately, by the construction of T_n, we don't have in general that $v \in \mathrm{T}_n$ implies that $v \in \mathrm{T}_{n'}$, see the example of Figure 3.3.

However, it is still true, as $1 + a > 0$, that:

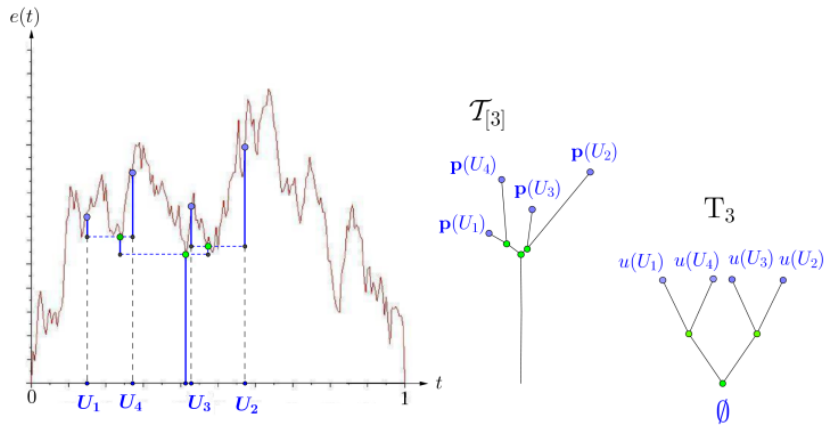
$$\sum_{v \in \mathrm{T}_n} |\mathrm{T}_{n,v}|^{1+a} \leq \sum_{v' \in \mathrm{T}_{n'}} |\mathrm{T}_{n',v'}|^{1+a}. \tag{3.38}$$

Let $n \in \mathbb{N}^*$. There exists a unique $n' \in \mathbb{N}^*$ such that $\varphi(n') \leq n < \varphi(n'+1)$. We obtain from (3.38) that:

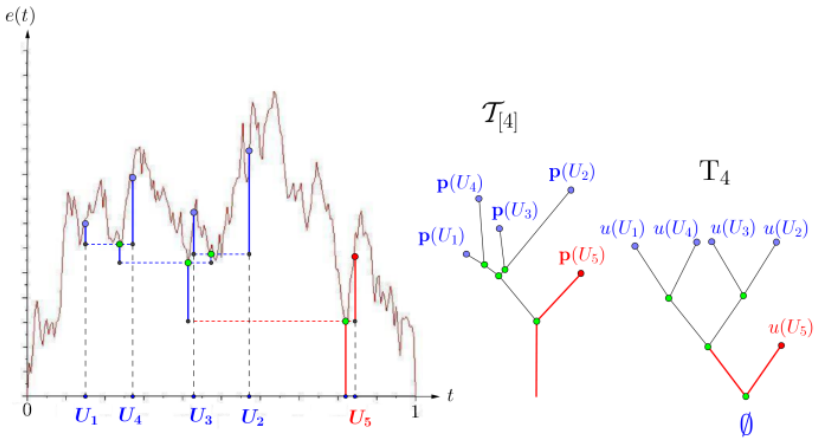
$$\left(\frac{2\varphi(n') + 1}{2\varphi(n'+1) + 1} \right)^{a+\frac{3}{2}} A_{\varphi(n')}(x^a) \leq A_n(x^a) \leq \left(\frac{2\varphi(n'+1) + 1}{2\varphi(n') + 1} \right)^{a+\frac{3}{2}} A_{\varphi(n'+1)}(x^a).$$

As $\lim_{n'\to+\infty} \varphi(n')/\varphi(n'+1) = 1$, we deduce that a.s. $\lim_{n\to+\infty} A_n(x^a) = \sqrt{2\alpha}\ \Phi_e(x^a)$.

In particular, for all $a \in (-1/2, 0]$, a.s. for all $k \in \mathbb{N}$, we have $\lim_{n\to+\infty} A_n(x^{a+k}) = \sqrt{2\alpha}\ \Phi_e(x^{a+k})$. Since on $[0,1]$, the convergence of moments implies the weak convergence of measure, we deduce that a.s. the random measure $A_n(x^a\, \bullet)$ converges weakly towards $\sqrt{2\alpha}\ \Phi_e(x^a\, \bullet)$. By taking a dense subset of a in $(-1/2, 0]$ and using monotonicity, we deduce that a.s. for all $a \in (-1/2, 0]$ the random measure $A_n(x^a\, \bullet)$ converges weakly towards $\sqrt{2\alpha}\ \Phi_e(x^a\, \bullet)$. This ends the proof of Theorem 3.4.



(a) $n = 4$



(b) $n' = 5$

Figure 3.3 – $\mathcal{T}_{[3]} \subset \mathcal{T}_{[4]}$ but $\mathrm{T}_3 \not\subset \mathrm{T}_4$

3.7 Proof of Proposition 3.10

3.7.1 A preliminary convergence in distribution

Let $(E_v, v \in \mathcal{U})$ be independent exponential random variables with mean 1 and independent of e. Let $f \in \mathcal{C}([0,1])$. We set for $v \in \mathrm{T}_n$:

$$X_{n,v} = |\mathrm{T}_n|^{-5/4} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) \quad \text{and} \quad Z_n(f) = \sum_{v \in \mathrm{T}_n} X_{n,v}\,(E_v - 1). \tag{3.39}$$

We have the following lemma.

Lemma 3.30. *Let $f \in \mathcal{C}([0,1])$ be locally Lipschitz continuous on $(0,1]$ such that $\| x^a f' \|_{esssup}$ is finite for some $a \in (0,1)$. We have the following convergence in distribution:*

$$(Z_n(f), A_n) \xrightarrow[n\to+\infty]{(d)} \left((2\alpha)^{1/4} \sqrt{\Phi_e(xf^2)}\; G, \sqrt{2\alpha}\; \Phi_e\right), \tag{3.40}$$

where G is a standard Gaussian random variable independent of e.

Proof. Let $f \in \mathcal{C}([0,1])$. We first assume that f is non-negative. We compute the Laplace transform of $Z_n(f)$ conditionally on T_n. Let $\lambda > 0$. Elementary computations give:

$$\mathbb{E}\left[\mathrm{e}^{-\lambda Z_n(f)} \,|\mathrm{T}_n\right] = \mathrm{e}^{\lambda \sum_{v\in\mathrm{T}_n} X_{n,v}}\, \mathbb{E}\left[\mathrm{e}^{-\lambda \sum_{v\in\mathrm{T}_n} X_{n,v} E_v} \,|\mathrm{T}_n\right] = \mathrm{e}^{\sum_{v\in\mathrm{T}_n} (\lambda X_{n,v} - \log(1+\lambda X_{n,v}))}.$$

For $x \geq 0$, we have $\frac{x^2}{2} - \frac{x^3}{3} \leq x - \log(1+x) \leq \frac{x^2}{2}$. Thanks to Theorem 3.4, we have:

$$\sum_{v \in \mathrm{T}_n} X_{n,v}^2 = A_n(xf^2) \xrightarrow[n\to+\infty]{\text{a.s.}} \sqrt{2\alpha}\, \Phi_e(xf^2)$$

and

$$\sum_{v \in \mathrm{T}_n} X_{n,v}^3 = |\mathrm{T}_n|^{-1/4} A_n(x^2 f^3) \xrightarrow[n\to+\infty]{\text{a.s.}} 0.$$

We deduce that a.s. $\lim_{n\to+\infty} \mathbb{E}\left[\mathrm{e}^{-\lambda Z_n(f)} \,|\mathrm{T}_n\right] = \exp\left(\lambda^2 \sqrt{2\alpha}\, \Phi_e(xf^2)/2\right)$. Let $K > 0$, and consider the event $B_K = \bigcap_{n\in\mathbb{N}} \{A_n(xf^2) \leq K\}$. Since on B_K, the term $\mathbb{E}\left[\mathrm{e}^{-\lambda Z_n(f)} \,|\mathrm{T}_n\right]$ is bounded by $\exp(\lambda^2 K/2)$, we deduce from dominated convergence that for any continuous bounded function g on the set of finite measure on $[0,1]$ (endowed with the topology of weak convergence), we have:

$$\begin{aligned}
\lim_{n\to+\infty} \mathbb{E}\left[\mathrm{e}^{-\lambda Z_n(f)}\, g(A_n) \mathbf{1}_{B_K}\right] &= \lim_{n\to+\infty} \mathbb{E}\left[\mathbb{E}\left[\mathrm{e}^{-\lambda Z_n(f)} \,|\mathrm{T}_n\right] g(A_n) \mathbf{1}_{B_K}\right] \\
&= \mathbb{E}\left[\mathrm{e}^{\lambda^2 \sqrt{2\alpha}\, \Phi_e(xf^2)/2}\, g(\sqrt{2\alpha}\, \Phi_e) \mathbf{1}_{B_K}\right] \\
&= \mathbb{E}\left[\mathrm{e}^{-\lambda (2\alpha)^{1/4} \sqrt{\Phi_e(xf^2)}\, G}\, g(\sqrt{2\alpha}\, \Phi_e) \mathbf{1}_{B_K}\right],
\end{aligned}$$

where G is a standard Gaussian random variable independent of e. We deduce that the convergence in distribution (3.40) holds conditionally on B_K. Since $A_n(xf^2)$ is finite for every n and converges a.s. to a finite limit, we get that for any $\varepsilon > 0$, there exists K_ε finite such that $\mathbb{P}(B_{K_\varepsilon}) \geq 1 - \varepsilon$. Then use Lemma 3.31 below to conclude that (3.40) holds for f non-negative.

In the general case, we set $f_+ = \max(0, f)$ and $f_- = \max(0, -f)$ so that $f = f_+ - f_-$. Notice that f_+ and f_- are non-negative and continuous. We have proved that (3.40) holds with f replaced by $\lambda_+ f_+ + \lambda_- f_-$ for any $\lambda_+ \geq 0$ and $\lambda_- \geq 0$. Since $f_+ f_- = 0$, this implies the following convergence in distribution:

$$(Z_n(f_+), Z_n(f_-), A_n) \xrightarrow[n\to+\infty]{(d)} \left((2\alpha)^{1/4} \sqrt{\Phi_e(xf_+^2)}\, G_+, (2\alpha)^{1/4} \sqrt{\Phi_e(xf_-^2)}\, G_-, \sqrt{2\alpha}\, \Phi_e\right),$$

where G_+ and G_- are independent standard Gaussian random variables independent of e. Then, using again that $f_+ f_- = 0$, we obtain that, conditionally on e, $\sqrt{\Phi_e(xf_+^2)}\, G_+ - \sqrt{\Phi_e(xf_-^2)}\, G_-$ is distributed as $\sqrt{\Phi_e(xf^2)}\, G$, where G is a standard Gaussian random variable independent of e. We deduce that (3.40) holds. This ends the proof. □

Lemma 3.31. *Let $(\Gamma_\varepsilon, \varepsilon > 0)$ be a sequence of events such that $\lim_{\varepsilon\to 0} \mathbb{P}(\Gamma_\varepsilon) = 1$. Let $(W_n, n \in \mathbb{N})$ and W be random variables taking values in a Polish space $\mathcal{M}$. Assume that for all $\varepsilon > 0$, conditionally on Γ_ε, the sequence $(W_n, n \in \mathbb{N})$ converges in distribution towards W. Then $(W_n, n \in \mathbb{N})$ converges in distribution towards W.*

Proof. Let g be a real-valued bounded continuous function defined on $\mathcal{M}$. It is enough to prove that $\lim_{n\to+\infty} |\mathbb{E}[g(W_n)] - \mathbb{E}[g(W)]| = 0$. By hypothesis, we have that for all $\varepsilon > 0$:

$$\lim_{n\to+\infty} \mathbb{E}[g(W_n)|\Gamma_\varepsilon] = \mathbb{E}[g(W)|\Gamma_\varepsilon].$$

We get:

$$|\mathbb{E}[g(W_n)] - \mathbb{E}[g(W)]| \leq |\mathbb{E}[g(W_n)|\Gamma_\varepsilon] - \mathbb{E}[g(W)|\Gamma_\varepsilon]| \mathbb{P}(\Gamma_\varepsilon) + 2 \,\|g\|_\infty\, \mathbb{P}(\Gamma_\varepsilon^c)$$

We deduce that $\limsup_{n\to+\infty} |\mathbb{E}[g(W_n)] - \mathbb{E}[g(W)]| \leq 2 \,\|g\|_\infty\, \mathbb{P}(\Gamma_\varepsilon^c)$. Since $\lim_{\varepsilon\to 0} \mathbb{P}(\Gamma_\varepsilon^c) = 0$, we deduce that $\lim_{n\to+\infty} |\mathbb{E}[g(W_n)] - \mathbb{E}[g(W)]| = 0$. This ends the proof. □

3.7.2 Proof of Proposition 3.10

We deduce Proposition 3.10 directly from Lemmas 3.32 and 3.33 below.
Using notations from Section 3.5, we set:

$$\Delta_n = \frac{1}{2\sqrt{\alpha}}\,|\mathrm{T}_n|^{1/4}(A_{1,n} - A_{2,n}).$$

Lemma 3.32. *Let $f \in \mathcal{C}([0,1])$ be locally Lipschitz continuous on $(0,1]$ such that $\| x^a f' \|_{esssup}$ is finite for some $a \in (0,1)$. We have the following convergence in probability:*

$$|\mathrm{T}_n|^{1/4}(A_n - \sqrt{2\alpha}\,\Phi_e)(f) - 2\sqrt{\alpha}\,\Delta_n(f) \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

Proof. We keep notations from Section 3.5. We have:

$$\Big||\mathrm{T}_n|^{1/4}(A_n - \sqrt{2\alpha}\,\Phi_e)(f) - 2\sqrt{\alpha}\,\Delta_n(f)\Big| \le \Delta_{1,n} + \Delta_{3,n} + \Delta_{4,n} + \Delta_{5,n},$$

where

$$\Delta_{1,n} = |\mathrm{T}_n|^{1/4}|A_n(f) - A_{1,n}(f)|, \quad \Delta_{3,n} = |\mathrm{T}_n|^{1/4}|A_{2,n}(f) - A_{3,n}(f)|,$$
$$\Delta_{4,n} = |\mathrm{T}_n|^{1/4}|A_{3,n}(f) - A_{4,n}(f)|, \quad \Delta_{5,n} = |\mathrm{T}_n|^{1/4}|A_{4,n}(f) - \sqrt{2\alpha}\,\Phi_e(f)|.$$

Using Lemmas 3.24, 3.26 and 3.27 part (i), we deduce the following convergence in probability:

$$\Delta_{1,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0, \quad \Delta_{3,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \quad \text{and} \quad \Delta_{4,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

We study the convergence of $\Delta_{5,n}$. We set:

$$I_n = \frac{1}{n+1}\sum_{k=1}^{n+1}\int_0^{e(U_k)} dr\, f(\sigma_{r,U_k}) - \int_0^1 ds \int_0^{e(s)} dr\, f(\sigma_{r,s}).$$

By conditioning with respect to e, we deduce that:

$$\mathbb{E}[I_n^2] \le \frac{1}{n+1}\,\mathbb{E}\left[\left(\int_0^{e(U_1)} dr\, f(\sigma_{r,U_1})\right)^2\right] \le \frac{\|f\|_\infty^2}{n+1}\,\mathbb{E}\Big[\int_0^1 ds\, e(s)^2\Big]. \tag{3.41}$$

Using the definition of $A_{4,n}(f)$, we get $\Delta_{5,n} \le \Delta_{6,n} + \sqrt{2\alpha}\,\Delta_{7,n}$ with

$$\Delta_{6,n} = |\mathrm{T}_n|^{1/4}\left|1 - \frac{|\mathrm{T}_n|^{3/2}}{2(n+1)\sqrt{2n}}\right| A_{4,n}(|f|) \quad \text{and} \quad \Delta_{7,n} = |\mathrm{T}_n|^{1/4}|I_n|.$$

From the a.s. convergence of $A_{4,n}(|f|)$ towards a finite limit, see Lemma 3.29, we deduce that a.s. $\lim_{n\to+\infty}\Delta_{6,n} = 0$. Since $\mathbb{E}\Big[\int_0^1 ds\, e(s)^2\Big]$ is finite, see [169], we deduce from (3.41) that $\lim_{n\to+\infty}\mathbb{E}[\Delta_{7,n}^2] = 0$. We obtain that:

$$\Delta_{5,n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

Then, we collect all the convergences together to get the result. □

Now, we study the convergence in distribution of $\Delta_n(f)$.

Lemma 3.33. *Let $f \in \mathcal{C}([0,1])$ be locally Lipschitz continuous on $(0,1]$ such that $\| x^a f' \|_{esssup}$ is finite for some $a \in (0,1)$. We have the following convergence in distribution:*

$$(2\sqrt{\alpha}\,\Delta_n(f), A_n) \xrightarrow[n\to+\infty]{(d)} \left((2\alpha)^{1/4}\sqrt{\Phi_e(xf^2)}\; G, \sqrt{2\alpha}\;\Phi_e\right), \tag{3.42}$$

where G is a standard Gaussian random variable independent of e.

Proof. According to Lemma 3.23, we get that $(\Delta_n(f), A_n)$ is distributed as $(\Delta'_n(f), A_n)$ where:

$$\Delta'_n(f) = |\mathrm{T}_n|^{-5/4} \sum_{v\in \mathrm{T}_n} |\mathrm{T}_{n,v}| f\left(\frac{|\mathrm{T}_{n,v}|}{|\mathrm{T}_n|}\right) Y'_{n,v}, \quad \text{with} \quad Y'_{n,v} = \sqrt{n}\left(\mathbb{E}\left[\frac{L'_n E_v}{S_{\mathrm{T}_n}}\right] - \frac{L'_n E_v}{S_{\mathrm{T}_n}}\right),$$

and $S_{\mathbf{t}} = \sum_{v\in\mathbf{t}} E_v$ for $\mathbf{t} \in \mathbb{T}$, with L'_n a random variable distributed as L_n, and thus with density given by (3.52), independent of T_n and $(E_u, u \in \mathcal{U})$ independent exponential random variables with mean 1, independent of L'_n and T_n. So it is enough to prove (3.42) with Δ_n replaced by Δ'_n.

Recall the definition (3.39) of $Z_n(f)$. Since L'_n is independent of $(E_u, u \in \mathcal{U})$ and T_n, we get:

$$\Delta'_n(f) = \frac{\sqrt{n}}{\sqrt{|\mathrm{T}_n|}}(\kappa_{1,n} + \kappa_{2,n})A_n(f) - \sqrt{n}\,\frac{L'_n}{S_{\mathrm{T}_n}} Z_n(f)$$

with

$$\kappa_{1,n} = |\mathrm{T}_n|^{3/4}\,(\mathbb{E}[L'_n] - L'_n)\mathbb{E}\left[\frac{E_\emptyset}{S_{\mathrm{T}_n}}\right] \quad \text{and} \quad \kappa_{2,n} = |\mathrm{T}_n|^{3/4}\,L'_n\left(\mathbb{E}\left[\frac{E_\emptyset}{S_{\mathrm{T}_n}}\right] - \frac{1}{S_{\mathrm{T}_n}}\right).$$

Thanks to Corollary 3.37 with $\alpha = \gamma = 1$ and $\beta = 0$, we have that:

$$\mathbb{E}[E_\emptyset / S_{\mathrm{T}_n}] = \Gamma(2n+1)/\Gamma(2n+2) = 1/|\mathrm{T}_n|.$$

Using (3.54), we get:

$$\mathbb{E}[|\kappa_{1,n}|] \le |\mathrm{T}_n|^{3/4}\sqrt{\mathrm{Var}\,(L'_n)}\;\frac{\Gamma(2n+1)}{\Gamma(2n+2)} \le \frac{1}{\sqrt{\alpha}}\;\frac{1}{(2n+1)^{1/4}}.$$

We deduce that $\lim_{n\to+\infty}\kappa_{1,n} = 0$ in probability. Using (3.53) and Corollary 3.37 (three times), we get:

$$\begin{aligned}\mathbb{E}[\kappa_{2,n}^2] &= |\mathrm{T}_n|^{3/2}\frac{n+1}{\alpha}\left(\frac{\Gamma(2n+1)^2}{\Gamma(2n+2)^2} + \frac{\Gamma(2n-1)}{\Gamma(2n+1)} - 2\frac{\Gamma(2n+1)}{\Gamma(2n+2)}\frac{\Gamma(2n)}{\Gamma(2n+1)}\right)\\ &= \frac{|\mathrm{T}_n|^{3/2}\;(n+1)(2n+3)}{\alpha 2n(2n+1)^2(2n-1)}.\end{aligned}$$

We deduce that $\lim_{n\to+\infty}\kappa_{2,n} = 0$ in probability.

We deduce from the law of large numbers that $\lim_{n\to+\infty} S_{\mathrm{T}_n}/|\mathrm{T}_n| = 1$ in probability. According to [2], we have that a.s. $\lim_{n\to+\infty} L_n/\sqrt{n} = 1/\sqrt{\alpha}$. This implies the following convergence in probability $\lim_{n\to+\infty} L'_n/\sqrt{n} = 1/\sqrt{\alpha}$. We obtain that:

$$\sqrt{n}\,\frac{L'_n}{S_{\mathrm{T}_n}} \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{2\sqrt{\alpha}}.$$

We deduce that $(2\sqrt{\alpha}\,\Delta'_n(f), A_n)$ has the same limit in distribution as $(-Z_n(f), A_n)$ as n goes to infinity. Then use Lemma 3.30 to get that (3.42) holds with Δ_n replaced by Δ'_n. This ends the proof of the Lemma. □

3.8 Appendix

3.8.1 Upper bounds for moments of the cost functional

According to [89], for $\beta > \frac{1}{2}$ and $k \in \mathbb{N}^*$, there exists a finite constant $C_{k,\beta}$ such that for all $n \in \mathbb{N}^*$,

$$\mathbb{E}\left[\left(\sum_{v\in \mathrm{T}_n} |\mathrm{T}_{n,v}|^{\beta}\right)^k\right] \leq C_{k,\beta}\,|\mathrm{T}_n|^{k(\beta+\frac{1}{2})}. \tag{3.43}$$

(Notice that (3.43) is stated in [89] with $\mathrm{T}^*_{n,v} = \mathrm{T}_{n,v}\backslash\mathcal{L}(\mathrm{T}_{n,v})$ instead of $\mathrm{T}_{n,v}$; but using that $|\mathrm{T}_{n,v}| = 2|\mathrm{T}^*_{n,v}| + 1$ it is elementary to get (3.43).)

The following lemma, which plays a key role in the proofs of Lemmas 3.24 and 3.25, is a direct consequence of these upper bounds.

Lemma 3.34. *For all $a \in [0, 1/2)$ and $f \in \mathcal{B}([0,1])$, we have for $k \in \mathbb{N}^*$:*

$$\mathbb{E}\left[|A_n(f)|^k\right] \leq C_{k,1-a}\|x^a f\|^k_\infty, \tag{3.44}$$

$$\mathbb{E}\left[A_n(xf^2)\right] \leq C_{1,2-2a}\|x^a f\|^2_\infty. \tag{3.45}$$

Proof. Let $k \in \mathbb{N}^*$. Using (3.43), we have:

$$\mathbb{E}\left[|A_n(f)|^k\right] \leq |\mathrm{T}_n|^{-\frac{3}{2}k}\|x^a f\|^k_\infty\,\mathbb{E}\left[\left(\sum_{v\in \mathrm{T}_n}\frac{|\mathrm{T}_{n,v}|^{1-a}}{|\mathrm{T}_n|^{-a}}\right)^k\right] \leq C_{k,1-a}\|x^a f\|^k_\infty,$$

which gives (3.44). Moreover, we also have:

$$\mathbb{E}\left[A_n(xf^2)\right] \leq |\mathrm{T}_n|^{-\frac{3}{2}}\|x^a f\|^2_\infty\,\mathbb{E}\left[\sum_{v\in \mathrm{T}_n}\frac{|\mathrm{T}_{n,v}|^{2-2a}}{|\mathrm{T}_n|^{1-2a}}\right] \leq C_{1,2-2a}\|x^a f\|^2_\infty$$

and we get (3.45). □

3.8.2 A lemma for binomial random variables

We give a lemma used for the proof of Lemma 3.27.

Lemma 3.35. *Let X be a binomial random variable with parameter $(n,p) \in \mathbb{N}^* \times (0,1)$.*

(i) For $a \in (0,1]$, we have

$$\mathbb{E}\left[(2X+1)^{-a}\right] \leq \left(1 \wedge \frac{1}{p(n+1)}\right)^a.$$

(ii) Let $f \in \mathcal{C}((0,1])$ be locally Lipschitz continuous and $b \in (0,1)$. Then we have:

$$\mathbb{E}\left[\left|f\left(\frac{2X+1}{2n+1}\right) - f(p)\right|\right] \leq \frac{\|\,x^b f'\,\|_{esssup}}{1-b}\left(p^{-\frac{b}{2}} + p^{\frac{1}{2}-b}\right) n^{-1/2}.$$

Proof. We prove (i). Let $a \in (0,1]$. Let X be a binomial random variable with parameter (n,p). An elementary computation gives that:

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1-(1-p)^{n+1}}{p(n+1)}. \tag{3.46}$$

Using Jensen inequality and (3.46), we get

$$\mathbb{E}\left[\left(\frac{1}{2X+1}\right)^a\right] \le \mathbb{E}\left[\frac{1}{2X+1}\right]^a \le \mathbb{E}\left[\frac{1}{1+X}\right]^a \le \left(1 \wedge \frac{1}{p(n+1)}\right)^a.$$

We prove (ii). Let $b \in (0,1)$. We have $\left|f\left(\frac{2X+1}{2n+1}\right) - f(p)\right| \le \|x^b f'\|_{\text{esssup}} \left|\int_p^{\frac{2X+1}{2n+1}} x^{-b}dx\right|$ and thus

$$\left|f\left(\frac{2X+1}{2n+1}\right) - f(p)\right| \le \frac{\|x^b f'\|_{\text{esssup}}}{1-b}\left|\left(\frac{2X+1}{2n+1}\right)^{1-b} - p^{1-b}\right|. \tag{3.47}$$

We decompose the right-hand side term into two parts:

$$\left|\left(\frac{2X+1}{2n+1}\right)^{1-b} - p^{1-b}\right| \le \left|p^{1-b} - \left(\frac{X}{n}\right)^{1-b}\right| + \left|\left(\frac{2X+1}{2n+1}\right)^{1-b} - \left(\frac{X}{n}\right)^{1-b}\right|. \tag{3.48}$$

We shall use the following key inequality (consider first the case $x \ge y$ and then the case $x < y$): for all $x, y > 0$ and $0 < b < 1$, we have:

$$|x^{1-b} - y^{1-b}| \le x^{-b}|x-y|. \tag{3.49}$$

For the first term of the right hand side of (3.48), using (3.49), we have $\left|p^{1-b} - \left(\frac{X}{n}\right)^{1-b}\right| \le p^{-b}\left|p - \frac{X}{n}\right|$. Hence, we get:

$$\mathbb{E}\left[\left|p^{1-b} - \left(\frac{X}{n}\right)^{1-b}\right|\right] \le p^{-b}\sqrt{\text{Var}\ (X/n)} \le p^{\frac{1}{2}-b}n^{-1/2}. \tag{3.50}$$

For the second term of the right hand side of (3.48), using (3.49) again, we get:

$$\left|\left(\frac{2X+1}{2n+1}\right)^{1-b} - \left(\frac{X}{n}\right)^{1-b}\right| \le \left(\frac{2X+1}{2n+1}\right)^{-b}\left|\frac{2X+1}{2n+1} - \frac{X}{n}\right| \le \frac{(2n+1)^{b-1}}{(2X+1)^b}.$$

This gives, using (i) and $|1 \wedge (1/x)|^b \le x^{-b/2}$ for $x > 0$, that:

$$\mathbb{E}\left[\left|\left(\frac{2X+1}{2n+1}\right)^{1-b} - \left(\frac{X}{n}\right)^{1-b}\right|\right] \le (2n+1)^{b-1}p^{-\frac{b}{2}}(n+1)^{-\frac{b}{2}} \le p^{-\frac{b}{2}}n^{-1/2}. \tag{3.51}$$

Using (3.47), (3.48), (3.50) and (3.51), we get the expected result. □

3.8.3 Some results on the Gamma function

We give here some results on the moments of Gamma random variables.

Lemma 3.36. *Let $k, \ell, n \in (0, +\infty)$ and $\alpha, \beta, \gamma \in [0, +\infty)$ such that $k + \ell + n + \alpha + \beta > \gamma$. Let $\Gamma_k, \Gamma_\ell, \Gamma_n$ be three independent Gamma random variables with respective parameter $(k,1)$, $(\ell,1)$ and $(n,1)$. Then we have:*

$$\mathbb{E}\left[\frac{\Gamma_k^\alpha \Gamma_\ell^\beta}{(\Gamma_k + \Gamma_\ell + \Gamma_n)^\gamma}\right] = \frac{\Gamma(k+\alpha)}{\Gamma(k)}\frac{\Gamma(\ell+\beta)}{\Gamma(\ell)}\frac{\Gamma(k+\ell+n+\alpha+\beta-\gamma)}{\Gamma(k+\ell+n+\alpha+\beta)}.$$

Proof. Elementary computations give that for all non negative function $f \in \mathcal{B}([0,\infty))$,

$$\mathbb{E}\left[\Gamma_k^\alpha f\left(\Gamma_k\right)\right] = \mathbb{E}\left[\Gamma_k^\alpha\right]\mathbb{E}\left[f\left(\Gamma_{k+\alpha}\right)\right] = \frac{\Gamma(k+\alpha)}{\Gamma(k)}\mathbb{E}\left[f\left(\Gamma_{k+\alpha}\right)\right].$$

We deduce that:

$$\begin{aligned}\mathbb{E}\left[\frac{\Gamma_k^\alpha\,\Gamma_\ell^\beta}{(\Gamma_k+\Gamma_\ell+\Gamma_n)^\gamma}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\Gamma_k^\alpha\,\Gamma_\ell^\beta}{(\Gamma_k+\Gamma_\ell+\Gamma_n)^\gamma}\middle|\Gamma_\ell,\Gamma_n\right]\right]\\ &= \mathbb{E}\left[\Gamma_k^\alpha\right]\mathbb{E}\left[\frac{\Gamma_\ell^\beta}{\left(\Gamma_\ell+\tilde{\Gamma}_{k+n+\alpha}\right)^\gamma}\right]\\ &= \mathbb{E}\left[\Gamma_k^\alpha\right]\mathbb{E}\left[\Gamma_\ell^\beta\right]\mathbb{E}\left[\frac{1}{\Gamma^\gamma_{k+\ell+n+\alpha+\beta}}\right]\\ &= \frac{\Gamma(k+\alpha)}{\Gamma(k)}\frac{\Gamma(\ell+\beta)}{\Gamma(\ell)}\frac{\Gamma(k+\ell+n+\alpha+\beta-\gamma)}{\Gamma(k+\ell+n+\alpha+\beta)},\end{aligned}$$

where $\tilde{\Gamma}_{k+n+\alpha}$ is a Gamma random variable with parameter $(k+n+\alpha,1)$ independent of Γ_ℓ, and $\Gamma_{k+\ell+n+\alpha+\beta}$ is a Gamma random variable with parameter $(k+\ell+n+\alpha+\beta,1)$. □

We directly deduce the following result.

Corollary 3.37. *Let $m\geq 2$. Let $(E_i,\,1\leq i\leq m)$ be independent exponential random variables with parameter 1 and $S_m=\sum_{i=1}^m E_i$. Then for all $\alpha,\beta,\gamma\in[0,+\infty)$ such that $m+\alpha+\beta>\gamma$, we have*

$$\mathbb{E}\left[\frac{E_1^\alpha E_2^\beta}{S_m^\gamma}\right]=\Gamma(1+\alpha)\Gamma(1+\beta)\frac{\Gamma(m+\alpha+\beta-\gamma)}{\Gamma(m+\alpha+\beta)}.$$

3.8.4 Elementary computations on the branch length of $\mathcal{T}_{[n]}$

We keep notations from Section 3.5. Recall that the density of $(h_{n,v},v\in\mathrm{T}_n)$ is, conditionally on T_n, given by (3.29). Recall $L_n=\sum_{v\in\mathrm{T}_n}h_{n,v}$ denotes the total length of $\mathcal{T}_{[n]}$. From Aldous [10], Pitman [163] (see Theorem 7.9) or Duquesne and Le Gall [71] or by standard computations, we get that the density of L_n, conditionally on T_n, is given by:

$$f_{L_n}(x)=2\frac{\alpha^{n+1}}{n!}x^{2n+1}\,\mathrm{e}^{-\alpha x^2}\,\mathbf{1}_{\{x>0\}}. \tag{3.52}$$

In particular, the random variable L_n is independent of T_n. The first two moments of L_n are given by

$$\mathbb{E}[L_n]=\frac{1}{\sqrt{\alpha}}\frac{\Gamma(n+\frac{3}{2})}{\Gamma(n+1)}=\frac{n+1}{\sqrt{\alpha}}\frac{\Gamma(n+\frac{3}{2})}{\Gamma(n+2)}\quad\text{and}\quad\mathbb{E}[L_n^2]=\frac{n+1}{\alpha}. \tag{3.53}$$

According to [98], we have that $(n+1)^{s-1}\leq\Gamma(n+s)/\Gamma(n+1)\leq n^{s-1}$ for $n\in\mathbb{N}^*$ and $s\in[0,1]$. Hence, we obtain:

$$\frac{1}{\sqrt{\alpha}}\frac{n+1}{\sqrt{n+2}}\leq\mathbb{E}[L_n]\leq\frac{\sqrt{n+1}}{\sqrt{\alpha}}\quad\text{and}\quad\mathrm{Var}\,(L_n)\leq\frac{1}{\alpha}\cdot \tag{3.54}$$

Using that $L_n=\sum_{v\in\mathrm{T}_n}h_{n,v}$ and that, conditionally on T_n, the random variables $(h_{n,v},v\in\mathrm{T}_n)$ are exchangeable, we deduce that $\mathbb{E}[h_{n,\emptyset}]=\mathbb{E}[L_n]/(2n+1)$ and thus:

$$\frac{1}{2\sqrt{\alpha(n+1)}}\leq\mathbb{E}[h_{n,\emptyset}]=\frac{1}{2\sqrt{\alpha}}\frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+1)}\leq\frac{1}{2\sqrt{\alpha n}}\cdot \tag{3.55}$$

We finish by a result on the covariance of the branch lengths, used in Lemma 3.25. We define $Y_{n,v}=\sqrt{n}(\mathbb{E}[h_{n,v}]-h_{n,v})$ for $v\in\mathrm{T}_n$. Notice that $(Y_{n,v},v\in\mathrm{T}_n)$ has an exchangeable distribution conditionally on T_n.

Lemma 3.38. *Let $n \in \mathbb{N}^*$. We have:*

$$\left|\mathbb{E}[Y_{n,\emptyset}Y_{n,1}]\right| \leq \frac{1}{8\alpha n} \quad \text{and} \quad \mathbb{E}[Y_{n,\emptyset}^2] \leq \frac{1}{2\alpha}. \tag{3.56}$$

Proof. Using Lemma 3.23 and its notations, and Corollary 3.37 and (3.53), we have, with $\mathbf{t} \in \mathbb{T}$ full binary such that $|\mathbf{t}| = 2n+1$:

$$\mathbb{E}[h_{n,\emptyset}h_{n,1}] = \mathbb{E}\left[L_n^2\right]\mathbb{E}\left[\frac{E_\emptyset E_1}{S_{\mathbf{t}}^2}\right] = \frac{n+1}{\alpha}\frac{1}{2(n+1)(2n+1)} = \frac{1}{2\alpha}\frac{1}{2n+1},$$

and

$$\mathbb{E}[h_{n,\emptyset}^2] = \mathbb{E}\left[L_n^2\right]\mathbb{E}\left[\frac{E_\emptyset^2}{S_{\mathbf{t}}^2}\right] = \frac{n+1}{\alpha}\frac{1}{(n+1)(2n+1)} = \frac{1}{\alpha}\frac{1}{2n+1}.$$

The lemma is then a consequence of these equalities and (3.55). □

3.8.5 A deterministic representation formula

Lemma 3.39. *Let $h \in \mathcal{C}_+([0,1])$. We have that for all $a > 0$:*

$$2\int_0^1 ds \int_0^{h(s)} dr\, \sigma_{r,s}(h)^a = a(a+1)\int_{[0,1]^2} |s'-s|^{a-1}\, m_h(s,s')\, dsds'. \tag{3.57}$$

Proof. In this proof only, we shall write $m(s,t)$ and $\sigma_{r,s}$ respectively for $m_h(s,t)$ and $\sigma_{r,s}(h)$. Recall that $\sigma_{r,s} = \int_0^1 dt\, \mathbf{1}_{\{m(s,t)\geq r\}}$. We deduce that $\int_0^1 dt\, m(s,t) = \int_0^{h(s)} dr\, \sigma_{r,s}$ for every $s \in [0,1]$. Hence, the result is obvious for $a = 1$.

If $g \in \mathcal{B}([0,1])$ is a non negative function such that $x^2 g \in \mathcal{C}^2([0,1])$ or if $g = x^{a-1}$ for $a > 0$, we set:

$$I(g) = \int_0^1 ds \int_0^{h(s)} dr\, \sigma_{r,s}\, g(\sigma_{r,s}) \quad \text{and} \quad J(g) = \int_{0<s<t<1} ds\, dt\, [x^2 g]''(t-s)\, m(s,t).$$

We then have to prove that $J(g) = I(g)$ for $g = x^{a-1}$ for all $a > 0$. First of all, remark that if $f, g \in \mathcal{B}([0,1])$ are non negative functions such that $x^2 f, x^2 g \in \mathcal{C}^2([0,1]$, we have:

$$|I(g) - I(f)| \leq \int_0^1 ds \int_0^{h(s)} dr\, \sigma_{r,s}\, |g(\sigma_{s,r}) - f(\sigma_{s,r})| \leq \|g-f\|_\infty \|h\|_\infty \tag{3.58}$$

and

$$|J(g)-J(f)| \leq \|(x^2 g)'' - (x^2 f)''\|_\infty \int_{0<s<t<1} ds\, dt m(s,t) \leq \|(x^2 g)'' - (x^2 f)''\|_\infty \|h\|_\infty. \tag{3.59}$$

The proof of $J(g) = I(g)$ when $g = x^{a-1}$ is divided in 3 steps. First of all, we prove the result when $a \in \mathbb{N}^*$, which gives the equality when g is polynomial. Then we get the case when $g \in \mathcal{C}^2([0,1])$ by Bernstein's approximation. This gives the case $a \geq 3$. Finally, we give the result for $a \in (0,3)\backslash\{1,2\}$.

1st step

Let $g = x^{a-1}$ with $a \in \mathbb{N}^*$. We have:

$$\begin{aligned} I(x^{a-1}) &= \int_0^1 ds \int_0^{h(s)} dr \left(\int_0^1 dt\, \mathbf{1}_{\{m(s,t)\geq r\}} \right)^a \\ &= \int_{[0,1]^{a+1}} ds\, ds_1 \dots ds_a \left(\int_0^{h(s_1)} dr\, \mathbf{1}_{\{m(s_1,s_2)\geq r\}} \dots \mathbf{1}_{\{m(s_1,s_{a+1})\geq r\}} \right) \\ &= \int_{[0,1]^{a+1}} ds\, ds_1 \dots ds_a \left(\min(m(s,s_1), \dots, m(s,s_a)) \right) \\ &= \int_{[0,1]^{a+1}} ds_1 \dots ds_{a+1} \left(\min(m(s_1,s_2), \dots, m(s_1,s_{a+1})) \right) \\ &= \int_{[0,1]^{a+1}} ds_1 \dots ds_{a+1} \left(m\left(\min_{1\leq i\leq a+1} s_i, \max_{1\leq i\leq a+1} s_i \right) \right), \end{aligned}$$

where we used that $\bigcup_{i=2}^{a+1} [s_1, s_i] = [\min_{1\leq i\leq a+1} s_i, \max_{1\leq i\leq a+1} s_i]$ for the last equality. By choosing $s = \min_{1\leq i\leq a+1} s_i$ and $t = \max_{1\leq i\leq a+1} s_i$, we have:

$$\begin{aligned} &\int_{[0,1]^{a+1}} ds_1 \dots ds_{a+1} \left(m\left(\min_{1\leq i\leq a+1} s_i, \max_{1\leq i\leq a+1} s_i \right) \right) \\ &\qquad = a(a+1) \int_{0<s<t<1} ds\, dt \int_{[s,t]^{a-1}} ds_1 \dots ds_{a-1}\, m(s,t) \\ &\qquad = a(a+1) \int_{0<s<t<1} m(s,t)(t-s)^{a-1}\, ds\, dt. \end{aligned} \tag{3.60}$$

This gives $I(x^{a-1}) = J(x^{a-1})$.

2nd step

Let $g \in \mathcal{C}^2([0,1])$ be a non negative function. For $n \in \mathbb{N}$, we define the associated Bernstein polynomial $B_n(g)$ by:

$$B_n(g)(x) = \sum_{k=0}^{n} \binom{n}{k} g(k/n)\, x^k (1-x)^{n-k}, \quad x \in [0,1].$$

It is well known (see for instance, Theorem 6.3.2 in [52]) that for every $k \in \mathbb{N}$ and for every $f \in \mathcal{C}^k([0,1])$, $\lim_{n\to\infty} \|f^{(k)} - B_n^{(k)}(f)\|_\infty = 0$. Using that $\|(x^2 B_n(g))'' - (x^2 g)''\|_\infty \leq 2\|B_n(g) - g\|_\infty + 4\|B_n'(g) - g'\|_\infty + \|B_n''(g) - g''\|_\infty$, we deduce from (3.58) and (3.59) that $J(g) = I(g)$.

3rd step

Let $g = x^{a-1}$ with $a \in (0,3)\backslash\{1,2\}$. We approximate g by functions in $\mathcal{C}^2([0,1])$. For $\delta \in (0,1)$, we define:

$$g_\delta(x) = \begin{cases} P_\delta(x) & \text{if } 0 \leq x \leq \delta, \\ g(x) & \text{if } \delta \leq x \leq 1, \end{cases}$$

where P_δ is the polynomial with degree 2 such that $P_\delta(\delta) = g(\delta) = \delta^{a-1}$, $P_\delta'(\delta) = g'(\delta) = (a-1)\delta^{a-2}$ and $P_\delta''(\delta) = g''(\delta) = (a-1)(a-2)\delta^{a-3}$.

We shall prove that $\lim_{\delta\to 0} I(g_\delta) = I(g)$. We have:

$$g_\delta''(x) = \begin{cases} g''(\delta) & \text{if } 0 \leq x \leq \delta, \\ g''(x) & \text{if } \delta \leq x \leq 1. \end{cases}$$

- Assume $a \in (0,1)$. Let $h = g_\gamma - g_\delta$ with $\delta, \gamma \in (0,1)$ such that $\delta < \gamma$. It is easy to check that $h'' \leq 0$ on $[0,1]$. Since $h'(1) = h(1) = 0$ by construction, we deduce that $h \leq 0$ on $[0,1]$. Hence, when δ tends to 0, the sequence $(g_\delta, 0 < \delta < 1)$ is non decreasing and converges on $(0,1]$ towards g. By monotone convergence theorem, we get $\lim_{\delta \to 0} I(g_\delta) = I(g)$.
- Assume $a \in (1,3)$. Notice that $(g_\delta, 0 < \delta < 1)$ is uniformly bounded by a constant. Hence, by dominated convergence theorem, we obtain that $\lim_{\delta \to 0} I(g_\delta) = I(g)$.

We now prove that $\lim_{\delta \to 0} J(g_\delta) = J(g)$. Remark that if $x \in (\delta, 1]$, $(x^2 g_\delta(x))'' = (x^2 g(x))''$, and that there exists a constant $C(a)$, which does not depend on δ, such that for all $x \in (0, \delta]$, we have $|(x^2 g_\delta(x))''| \leq C(a)\delta^{a-1}$. We get that:

$$\begin{aligned} |J(g_\delta) - J(g)| &= \left| \int_{0<s<t<1} m(s,t) \left((x^2 g_\delta(x))'' - (x^2 g(x))'' \right)_{x=t-s} ds dt \right| \\ &\leq \| h \|_\infty \int_0^\delta \left(\left| (x^2 g_\delta(x))'' - (x^2 g(x))'' \right| \right)_{x=r} dr \\ &\leq \| h \|_\infty \int_0^\delta \left(a(a+1) r^{a-1} + C(a)\delta^{a-1} \right) dr. \end{aligned}$$

We deduce that $\lim_{\delta \to 0} J(g_\delta) = J(g)$. Thanks to the 2nd step, we have $J(g_\delta) = I(g_\delta)$ for all $\delta \in (0,1)$. Letting δ goes down to 0, we deduce that $J(g) = I(g)$. □

3.8.6 Proof of the first part of Lemma 3.17 (finiteness of Z_β^H and (3.23))

We use the setting of [70] on Lévy trees. Let H be the height function of a stable Lévy tree with branching mechanism $\psi(\lambda) = \kappa \lambda^\gamma$, with $\gamma \in (1,2]$ and $\kappa > 0$.

Let $\mathbb{N}$ be the excursion measure of the height process and set $\sigma = \inf\{s > 0,\, H(s) = 0\}$ for the duration of the excursion so that: $\mathbb{N}[1 - \mathrm{e}^{-\lambda\sigma}] = \psi^{-1}(\lambda)$ for all $\lambda > 0$. According to Chapter VII in [21], ψ^{-1} is the Laplace exponent of a subordinator whose Lévy measure is denoted by π_*. Thus, the distribution of σ under $\mathbb{N}$ is π_* given by:

$$\pi_*(da) = \frac{1}{\gamma \kappa^{1/\gamma} \Gamma((\gamma-1)/\gamma)} \frac{da}{a^{1+\frac{1}{\gamma}}}.$$

Let $\mathbb{N}^{(a)}[\bullet] = \mathbb{N}[\bullet | \sigma = a]$ be the distribution of the excursion of the height process with duration a, so that:

$$\mathbb{N}[\bullet] = \int_0^\infty \pi_*(da)\, \mathbb{N}^{(a)}[\bullet].$$

In particular, we shall prove the result of Lemma 3.17 under $\mathbb{N}^{(1)}$. In this proof only, we shall write m for m_H defined by (3.13). We extend the definitions (3.14) and (3.15) as follows:

$$\sigma_{r,s} = \int_0^\sigma dt\, \mathbf{1}_{\{\min(s,t) \geq r\}} \quad \text{and} \quad Z_\beta^H = \int_0^\sigma ds \int_0^{H(s)} dr\, \sigma_{r,s}^{\beta-1} \quad \text{for } \beta > 0.$$

The integral in ds/σ in Z_β^H corresponds to taking a leaf at random in the Lévy tree. Using Bismut's decomposition of the Lévy tree, see Theorem 4.5 in [71] or Theorem 2.1 in [1], we get that, since $\psi'(0) = 0$, then under $\mathbb{N}[\sigma \bullet]$, the height $H(U)$, with U uniformly distributed over $[0,\sigma]$, is "distributed" as $\mathcal{H}$ with Lebesgue "distribution" on $(0, +\infty)$. It also implies that under $\mathbb{N}[\sigma\bullet]$, the random variable $\big(H(U), (\sigma_{H(U)-r,U}, r \in [0, H(U)])\big)$ is "distributed" as $(\mathcal{H}, (S_t, t \in [0,\mathcal{H}]))$, where $S = (S_t, t \geq 0)$ is a subordinator, with Laplace exponent say ϕ, independent of $\mathcal{H}$.

We prove (3.23) and get as a direct consequence using monotonicity, that $\mathbb{N}^{(1)}$-a.s., for all $\beta > 1/\gamma$, Z_β^H is finite. Using that:

$$(\psi^{-1})'(\lambda) = \mathbb{N}\left[\sigma\, \mathrm{e}^{-\lambda\sigma}\right] = \mathbb{N}\left[\sigma\, \mathrm{e}^{-\lambda\sigma_{0,U}}\right] = \mathbb{E}\left[\mathrm{e}^{-\lambda S_{\mathcal{H}}}\right] = \int_0^\infty dt\, \mathbb{E}\left[\mathrm{e}^{-\lambda S_t}\right] = \frac{1}{\phi(\lambda)}, \tag{3.61}$$

we deduce that:

$$\phi(\lambda) = \frac{1}{(\psi^{-1})'(\lambda)} = \gamma\kappa^{1/\gamma}\lambda^{(\gamma-1)/\gamma}. \tag{3.62}$$

Notice in particular that S_t is distributed as $t^{\gamma/(\gamma-1)}S_1$. We shall need later in the proof the following computation:

$$\mathbb{E}\left[S_1^{-(\gamma-1)/\gamma}\right] = \frac{1}{\Gamma\left(\frac{\gamma-1}{\gamma}\right)}\int_0^\infty dt\, t^{-1/\gamma}\mathbb{E}\left[\mathrm{e}^{-tS_1}\right] = \frac{1}{\kappa^{1/\gamma}(\gamma-1)\Gamma\left(\frac{\gamma-1}{\gamma}\right)}\cdot \tag{3.63}$$

We set $\Lambda(\lambda) = \mathbb{N}\left[Z_\beta^H\, \mathrm{e}^{-\lambda\sigma}\right]$ for $\lambda > 0$. Using Bismut's decomposition again, we get:

$$\begin{aligned}\Lambda(\lambda) = \mathbb{N}\left[\sigma\int_0^{H(U)} dr\, \sigma_{H(U)-r,U}^{\beta-1}\, \mathrm{e}^{-\lambda\sigma_{0,U}}\right] &= \mathbb{E}\left[\int_0^{\mathcal{H}} dr\, S_r^{\beta-1}\, \mathrm{e}^{-\lambda S_{\mathcal{H}}}\right]\\ &= \mathbb{E}\left[\int_0^\infty dt\int_0^t dr\, S_r^{\beta-1}\, \mathrm{e}^{-\lambda S_t}\right]\\ &= \mathbb{E}\left[\int_0^\infty dt\int_0^\infty dr\, S_r^{\beta-1}\, \mathrm{e}^{-\lambda S_{t+r}}\right].\end{aligned}$$

We have:

$$\begin{aligned}\Lambda(\lambda) &= \mathbb{E}\left[\int_0^\infty dt\, \mathrm{e}^{-\lambda S_t}\right]\mathbb{E}\left[\int_0^\infty dr\, S_r^{\beta-1}\, \mathrm{e}^{-\lambda S_r}\right]\\ &= \frac{1}{\phi(\lambda)}\,\mathbb{E}\left[S_1^{\beta-1}\int_0^\infty dr\, r^{(\beta-1)\gamma/(\gamma-1)}\, \mathrm{e}^{-\lambda r^{\gamma/(\gamma-1)}S_1}\right]\\ &= \frac{1}{\phi(\lambda)}\,\mathbb{E}\left[S_1^{-(\gamma-1)/\gamma}\right]\lambda^{-\beta+(1/\gamma)}\,\frac{\gamma-1}{\gamma}\int_0^\infty du\, u^{\beta-1-(1/\gamma)}\, \mathrm{e}^{-u},\end{aligned}$$

where we used that S has stationary independent increments for the first equality, (3.61) and that S_r is distributed as $r^{\gamma/(\gamma-1)}S_1$ for the second, and the change of variable $u = \lambda S_1 r^{\gamma/(\gamma-1)}$ for the last. Then use (3.63) and (3.62) to deduce that:

$$\Lambda(\lambda) = \frac{\Gamma\left(\beta - \frac{1}{\gamma}\right)}{\gamma^2\kappa^{2/\gamma}\Gamma\left(\frac{\gamma-1}{\gamma}\right)}\lambda^{-1-\beta+\frac{2}{\gamma}}. \tag{3.64}$$

On the other hand, we set $G(a) = \mathbb{N}^{(a)}[Z_\beta^H]$ so that:

$$\Lambda(\lambda) = \int_0^\infty \pi_*(da)\, G(a)\, \mathrm{e}^{-\lambda a}\,.$$

We deduce from the scaling property of the height function that, under $\mathbb{N}^{(a)}$, the random variable $\Big((H(s), s\in[0,a]), (\sigma_{r,s}; r\in[0,H(s)], s\in[0,a])\Big)$ is distributed as the random variable $\Big((a^{(\gamma-1)/\gamma}H(s/a),\ s\in[0,a]), (a\sigma_{r,s/a}; r\in[0,a^{(\gamma-1)/\gamma}H(s/a)], s\in[0,a])\Big)$ under

$\mathbb{N}^{(1)}$. This implies that Z_β^H is under $\mathbb{N}^{(a)}$ distributed as $a^{\beta+1-1/\gamma} Z_\beta^H$ under $\mathbb{N}^{(1)}$. This gives $G(a) = a^{\beta+1-1/\gamma} G(1)$. We deduce that:

$$\Lambda(\lambda) = G(1) \int_0^\infty \pi_*(da)\, a^{\beta+1-\frac{1}{\gamma}}\, \mathrm{e}^{-\lambda a} = G(1) \frac{\Gamma\left(\beta+1-\frac{2}{\gamma}\right)}{\gamma \kappa^{1/\gamma} \Gamma\left(\frac{\gamma-1}{\gamma}\right)} \lambda^{-\beta-1+\frac{2}{\gamma}}.$$

Then use (3.64) to get that for all $\beta > 0$:

$$\mathbb{N}^{(1)}[Z_\beta^H] = G(1) = \frac{1}{\gamma \kappa^{1/\gamma}} \frac{\Gamma\left(\beta-\frac{1}{\gamma}\right)}{\Gamma\left(\beta+1-\frac{2}{\gamma}\right)}.$$

This gives (3.23) and that $\mathbb{N}^{(1)}$-a.s., for all $\beta > 1/\gamma$, Z_β^H is finite.

We prove now that $\mathbb{N}^{(1)}$-a.s., for all $\beta \in (0, 1/\gamma]$, Z_β^H is infinite. Let $\beta \in (0, 1/\gamma]$. Let U be uniform on $[0, \sigma]$ under $\mathbb{N}$. According to the first part of the proof, we deduce from the Bismut's decomposition that $\int_0^{H(U)} dr\ \sigma_{r,U}^{\beta-1}$ is, under $\mathbb{N}[\sigma \bullet |H(U) = t]$, distributed as $\int_0^t dr\, S_r^{\beta-1}$. Thanks to [21] see Theorem 11 in chapter III and since S is a stable subordinator with index $(\gamma-1)/\gamma$, we have that $\limsup_{r\to 0+} S_r/h(r) > 0$ a.s. for $h(r) = r^{\gamma/(\gamma-1)} \log(|\log(r)|)^{-1/(\gamma-1)}$. As $\beta \in (0, 1/\gamma]$, we have $\int_0 dr\, h(r)^{\beta-1} = +\infty$. This implies that a.s. $\int_0 dr\, S_r^{\beta-1} = +\infty$. We deduce that $\mathbb{N}$-a.e. ds-a.e. on $[0, \sigma]$, $\int_0^{H(s)} dr\ \sigma_{r,s}^{\beta-1} = +\infty$. This gives that $\mathbb{N}$-a.e. $Z_\beta^H = +\infty$. Then use the scaling to deduce that $\mathbb{N}^{(1)}$-a.s. $Z_\beta^H = +\infty$.

Acknowledgments. We would like to thank the Referees and the Associate Editor for their useful comments which helped to improve the paper. We also thank T. Duquesne for discussions on the height process of Lévy trees.

CHAPITRE 4

ASYMPTOTIQUES POUR LA FONCTION DE RÉPARTITION EMPIRIQUE DES DEGRÉS ET LES DENSITÉS D'HOMOMORPHISMES DE GRAPHES ALÉATOIRES ÉCHANTILLONNÉS À PARTIR D'UN GRAPHON

Version légèrement modifiée de l'article [53]
Asymptotics for the cumulative distribution of the degrees and homomorphism densities for random graphs sampled from a graphon
soumis pour publication.

Abstract. We give asymptotics for the cumulative distribution function (CDF) for degrees of large dense random graphs sampled from a graphon. The proof is based on precise asymptotics for binomial random variables.

Replacing the indicator function in the empirical CDF by a smoother function, we get general asymptotic results for functionals of homomorphism densities for partially labeled graphs with smoother functions. This general setting allows to recover recent results on asymptotics for homomorphism densities of sampled graphon.

4.1 Introduction

The Internet, social networks or biological networks can be represented by large random graphs. Understanding their structure is an important issue in Mathematics. The degree sequence is one of the key objects used to get informations about graphs. The degree sequences of real world networks have attracted a lot of attention during the last years because their distributions are significantly different from the degree distributions studied in the classical models of random graphs such as the Erdős-Rényi model where the degree distribution is approximately Poisson when the number of nodes is large. They followed a power-law distribution, see for instance, Newmann [158], Chung et al [49], Diaconis and Blitzstein [26] and Newman, Barabasi and Watts [155]. See also Molloy and Reed [147, 148] and Newman, Strogatz and Watts [156] in the framework of sparse graphs.

In this paper, we shall consider the cumulative distribution function (CDF) of degrees of large dense random graphs sampled from a graphon, extending results from Bickel, Chen and Levina [25]. The theory of graphon or limits of sequence of dense graphs was developed by Lovász and Szegedy [139] and Borg, Chayes, Lovász, Sós and Vesztergombi [40]. The asymptotics on the empirical CDF of degrees, see the theorem in Section 4.1.1, could be used to test

if a large dense graph is sampled from a given graphon. This result is a first step for giving a non-parametric test for identifying the degree function of a large random graph in the spirit of the Kolmogorov-Smirnov test for the equality of probability distribution from a sample of independent identically distributed random variables.

If we replace the indicator function in the empirical CDF by a smoother function, we get general results on the fluctuations for functionals of homomorphism densities for partially labeled graphs. As an application, when considering homomorphism densities for sampled graphon, we recover results from Féray, Méliot and Nikeghbali [84].

4.1.1 Convergence of CDF of empirical degrees for large random graphs

We consider simple finite graphs, that is graphs without self-loops and multiple edges between any pair of vertices. We denote by $\mathcal{F}$ the set of all simple finite graphs.

There exists several equivalent notions of convergence for sequences of finite dense graphs (that is graphs where the number of edges is close to the maximal number of edges), for instance in terms of metric convergence (with the cut distance) or in terms of the convergence of subgraph densities, see [40] or Lovász [138].

When it exists, the limit of a sequence of dense graphs can be represented by a graphon i.e. a symmetric, measurable function $W : [0,1]^2 \to [0,1]$, up to a measure preserving bijection. A graphon W may be thought of as the weight matrix of an infinite graph whose set of vertices is the continuous unit interval, so that $W(x,y)$ represents the weight of the edge between vertices x and y.

Moreover, it is possible to sample simple graphs, with a given number of vertices, from a graphon W (called W-random graphs). Let $X = (X_i : i \in \mathbb{N}^*)$ be a sequence of independent random variables uniformly distributed on the interval $[0,1]$. To construct the W-random graph with vertices $[n] := \{1, \dots, n\}$, denoted by G_n, for each pair of distinct vertices $i \neq j$, elements of $[n]$, connect i and j with probability $W(X_i, X_j)$, independently of all other edges (see also Section 4.2.4). If needed, we shall stress the dependence on W and write $G_n(W)$ for G_n. By this construction, we get a sequence of random graphs $(G_n : n \in \mathbb{N}^*)$ which converges almost surely towards the graphon W, see for instance Proposition 11.32 in [138].

We define the degree function $D = (D(x) : x \in [0,1])$ of the graphon W by:

$$D(x) = \int_0^1 W(x,y)dy.$$

And we consider the empirical CDF $\Pi_n = (\Pi_n(y) : y \in [0,1])$ of the normalized degrees of the graph G_n defined by

$$\Pi_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{D_i^{(n)} \leq D(y)\right\}},$$

where $(n-1)D_i^{(n)}$ is the degree of the vertex i in G_n.

Bickel, Chen and Levina [25], Theorem 5 (with $m = 1$), proved the convergence in distribution and the convergence of the second moments of $\Pi_n(y)$ towards y. We improve the results given in [25]: under the condition that D is increasing[1] on $[0,1]$, we have the almost sure convergence of $\Pi_n(y)$ towards y, uniformly on $[0,1]$. This is a consequence of the more general result given by Theorem 4.9 (see Subsection 4.1.2 and Remark 4.22 for more details).

[1] Since the graphon is defined up to a measure preserving one-to-one map on $[0,1]$, there exists an equivalent version of the graphon for which the degree function is non-decreasing. If the degree function is increasing, then this version is unique in L^1 and this is the version which is considered in this section.

In a different direction, Chatterjee and Diaconis [45] considered the convergence of uniformly chosen random graphs with a given CDF of degrees towards an exponential graphon with given degree function.

We also get the fluctuations associated to the almost sure convergence of Π_n. If W satisfies some regularity conditions given by (4.62), which in particular imply that D is of class $\mathcal{C}^1$, then we have the following result on the convergence in distribution of finite-dimensional marginals for Π_n.

Theorem (Theorem 4.23)**.** *Assume that W satisfies condition (4.62). Then we have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1)\right) \xrightarrow[n\to+\infty]{(fdd)} \chi,$$

where $(\chi_y : y \in (0,1))$ is a centered Gaussian process defined, for all $y \in (0,1)$ by:

$$\chi_y = \int_0^1 (\rho(y,u) - \bar{\rho}(y))dB_u,$$

with $B = (B_u, u \geq 0)$ a standard Brownian motion, and $(\rho(y,u) : u \in [0,1])$ and $\bar{\rho}(y)$ defined for $y \in (0,1)$ by:

$$\rho(y,u) = \mathbf{1}_{[0,y]}(u) - \frac{W(y,u)}{D'(y)} \quad \text{and} \quad \bar{\rho}(y) = \int_0^1 \rho(y,u)du.$$

The covariance kernel $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$ of the Gaussian process χ is explicitly given by Equations (4.64), (4.65) and (4.66) which define respectively Σ_1, Σ_2 and Σ_3. In particular, we deduce that the variance of $\chi(y)$, for $y \in (0,1)$ is given by the elementary formula:

$$\Sigma(y,y) = y(1-y) + \frac{1}{D'(y)^2}\left(\int_0^1 W(y,x)^2dx - D(y)^2\right) + \frac{2}{D'(y)}\left(D(y)y - \int_0^y W(y,x)dx\right).$$

The proof of this result relies on uniform Edgeworth expansions for binomial random variables, see Bhattacharya and Rao [22], and Stein's method for binomial random vectors, see Bentkus [20]. The convergence of the process in the Skorokhod space could presumably be proved using similar but more involved arguments. More generally, following van der Vaart [188], Chapter 19 on convergence of empirical CDF of independent identically distributed random variables, if would be natural to study the uniform convergence of $\frac{1}{n}\sum_{i=1}^n f(D_i^{(n)})$ when f belongs to a certain class of functions.

The asymptotics on the CDF of empirical degrees appear formally as a limiting case of the asymptotics of $\frac{1}{n}\sum_{i=1}^n f(D_i^{(n)})$ with f smooth. This is developed in Section 4.1.3. We shall in fact adopt in this section a more general point of view as we replace the normalized degree sequence by a sequence of homomorphism densities for partially labeled graphs.

4.1.2 Convergence of sequence of dense graphs towards graphons

Recall that one of the equivalent notions of convergence of sequences of dense graphs is given by the convergence of subgraph densities. It is the latter one that will interest us. We first recall the notion of homomorphism densities. For two simple finite graphs F and G with respectively $v(F)$ and $v(G)$ vertices, let $\mathrm{Inj}(F,G)$ denote the set of injective homomorphisms (injective adjacency-preserving maps) from F to G (see Subsection 4.2.2 for

a precise definition). We define the injective homomorphism density from F to G by the following normalized quantity:

$$t_{\mathrm{inj}}(F,G) = \frac{|\mathrm{Inj}(F,G)|}{A_{v(G)}^{v(F)}},$$

where we have for all $n \geq k \geq 1$, $A_n^k = n!/(n-k)!$. In the same way, we can define the density of induced homomorphisms (which are injective homomorphisms that also preserve non-adjacency), see (4.24). Some authors study subgraph counts rather than homomorphism densities, but the two quantities are related, see Bollobás and Riordan [32], Section 2.1, so that results on homomorphism densities can be translated into results for subgraph counts.

A sequence of dense simple finite graphs $(H_n : n \in \mathbb{N}^*)$ is called convergent if the sequence $(t_{\mathrm{inj}}(F,H_n) : n \in \mathbb{N}^*)$ has a limit for every $F \in \mathcal{F}$. The limit can be represented by a graphon, say W and we have that for every $F \in \mathcal{F}$:

$$\lim_{n\to\infty} t_{\mathrm{inj}}(F,H_n) = t(F,W),$$

where

$$t(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i,x_j) \prod_{k\in V(F)} dx_k.$$

According to [138], Proposition 11.32, the sequence of W-random graphs $(G_n : n \in \mathbb{N}^*)$ converges a.s. towards W, that is for all $F \in \mathcal{F}$, a.s.:

$$\lim_{n\to\infty} t_{\mathrm{inj}}(F,G_n) = t(F,W). \tag{4.1}$$

In the Erdős-Rényi case, that is when $W \equiv \mathfrak{p}$ is constant, the fluctuations associated to this almost sure convergence are of order n: for all $F \in \mathcal{F}$ with p vertices and e edges, we have the following convergence in distribution:

$$n\left(t_{\mathrm{inj}}(F,G_n(\mathfrak{p})) - \mathfrak{p}^e\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, 2e^2\mathfrak{p}^{2e-1}(1-\mathfrak{p})\right),$$

where $\mathcal{N}(m,\sigma^2)$ denotes a Gaussian random variable with mean m and variance σ^2. There are several proofs of this central limit theorem. Nowicki [159] and Janson and Nowicki [118] used the theory of U-statistics to prove the asymptotic normality of subgraph counts and induced subgraph counts. They also obtained the asymptotic normality of vectors of subgraph counts and induced subgraph counts. In the particular case of the joint distribution of the count of edges, triangles and two-stars, Reinert and Röllin [167], Proposition 2, obtained bounds on the approximation. Using discrete Malliavin calculus, Krokowski and Thäle [132] generalized the result of [167] (in a different probability metric) and get the rate of convergence associated to the multivariate central limit theorem given in [118]. See also Féray, Méliot and Nikeghbali [83], Section 10, for the mod-Gaussian convergence of homomorphism densities.

The asymptotics of normalized subgraph counts have also been studied when the parameter $\mathfrak{p}$ of the Erdős-Rényi graphs depends on n, see for example Ruciński [174], Nowicki and Wierman [161], Barbour, Karoński and Ruciński [18], and Gilmer and Kopparty [100].

In the general framework of graphons, the speed of convergence in the invariance principle is of order $\sqrt{n}$, except for degenerate cases such as the Erdős-Rényi case. This result was given by Féray, Méliot and Nikeghbali [84], Theorem 21: for all $F \in \mathcal{F}$, we have the following convergence in distribution:

$$\sqrt{n}\left(t_{\mathrm{inj}}(F,G_n) - t(F,W)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0,\sigma(F)^2\right), \tag{4.2}$$

where, with $V(F)$ the set of vertices of F and $v(F)$ its cardinality,

$$\sigma(F)^2 = \sum_{q,q' \in V(F)} t\big((F \bowtie F)(q,q'), W\big) - v(F)^2\, t(F,W)^2$$

and $(F \bowtie F')(q,q')$ is the disjoint union of the two simple finite graphs F and F' where we identify the vertices $q \in F$ and $q' \in F'$ (see point (iii) of Remark 4.12, for more details). Notice that in the Erdős-Rényi case, that is when W is a constant graphon, the asymptotic variance $\sigma(F)^2$ is equal to 0, which is consistent with the previous paragraph since the speed is of order n.

Using Stein's method, Fang and Röllin [81] obtained the rate of convergence for the multivariate normal approximation of the joint distribution of the normalized edge count and the corrected and normalized 4-cycle count. As a consequence, they get a confidence interval to test if a given graph G comes from an Erdős-Rényi random graph model or a non constant graphon-random graph model. Maugis, Priebe, Olhede and Wolfe [143] gave a central limit theorem for subgraph counts observed in a network sample of W-random graphs drawn from the same graphon W when the number of observations in the sample increases but the number of vertices in each graph observation remains finite. They also get a central limit theorem in the case where all the graph observations may be generated from different graphons. This allows to test if the graph observations come from a specified model. When considering sequences of graphons which tend to 0, then there is a Poisson approximation of subgraph counts. In this direction, Coulson, Gaunt and Reinert [51], Corollary 4.1, used the Stein method to establish an effective Poisson approximation for the distribution of the number of subgraphs in the graphon model which are isomorphic to some fixed strictly balanced graph.

Motivated by those results, we present in the next section an invariance principle for the distribution of homomorphism densities of partially labeled graphs for W-random graphs which can be seen as a generalization of (4.2).

4.1.3 Asymptotics for homomorphism densities of partially labeled graphs for large random graphs

Let $n \in \mathbb{N}^*$ and $k \in [n]$. We define the set $\mathcal{S}_{n,k}$ of all $[n]$-words of length k such that all characters are distinct, see (4.7). Notice that $|\mathcal{S}_{n,k}| = A_n^k = n!/(n-k)!$.

We generalize homomorphism densities for partially labeled graphs. Let $F, G \in \mathcal{F}$ be two simple graphs with $V(F) = [p]$ and $V(G) = [n]$. Assume $n \geq p > k \geq 1$. Let $\ell \in \mathcal{S}_{p,k}$ and $\alpha \in \mathcal{S}_{n,k}$. We define $\mathrm{Inj}(F^\ell, G^\alpha)$ the set of injective homomorphisms f from F into G such that $\mathrm{f}(\ell_i) = \alpha_i$ for all $i \in [k]$, and its density:

$$t_{\mathrm{inj}}(F^\ell, G^\alpha) = \frac{|\mathrm{Inj}(F^\ell, G^\alpha)|}{A_{n-k}^{p-k}}.$$

We define the random probability measure $\Gamma_n^{F,\ell}$ on $([0,1], \mathcal{B}([0,1]))$, with $\mathcal{B}([0,1])$ the Borel σ-field on $[0,1]$, by: for all measurable non-negative function g defined on $[0,1]$,

$$\Gamma_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} g\left(t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right). \tag{4.3}$$

We prove, see Theorem 4.9, the almost sure convergence for the weak topology of the sequence $\left(\Gamma_n^{F,\ell}(dx) : n \in \mathbb{N}^*\right)$ of random probability measure on $[0,1]$ towards a deterministic probability measure $\Gamma^{F,\ell}(dx)$, see definition (4.42).

- If we take $g = \mathrm{Id}$ in (4.3), we recover the almost sure convergence given in (4.1) as according to (4.21):
$$t_{\mathrm{inj}}(F, G_n) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\mathrm{inj}}(F^\ell, G_n^\alpha).$$

- If we take $g = \mathbf{1}_{[0,D(y)]}$ with $y \in (0,1)$ and $F = K_2$ (where K_2 denotes the complete graph with two vertices) in (4.3) and using the expression of $\Gamma^{F,\ell}$ given in Remark 4.8, (ii), we have, with $\bullet$ any vertex of K_2, that:
$$\Gamma_n^{K_2,\bullet}(g) = \Pi_n(y) \quad \text{and} \quad \Gamma^{K_2,\bullet}(g) = y.$$
Then, by Theorem 4.9, under the condition that D is strictly increasing on $(0,1)$, we have the almost sure convergence of $\Pi_n(y)$ towards y, see Remark 4.22.

We also have the fluctuations associated to this almost sure convergence, see Theorem 4.11 for a multidimensional version.

Theorem. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}$ be a simple finite graphs with $V(F) = [p]$, $\ell \in \mathcal{M}_p$, with $k = |\ell|$. Then, for all $g \in \mathcal{C}^2([0,1])$, we have the following convergence in distribution:*
$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$
with $\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(\mathcal{U}_g^{F,\ell}\right)$ and $\mathcal{U}_g^{F,\ell}$ is defined in (4.43).

Notice $\sigma^{F,\ell}(g)^2$ is an integral involving g and g'. The asymptotic results are still true when we consider a family of $d \geq 1$ simple graphs $F = (F_m : 1 \leq m \leq d) \in \mathcal{F}^d$ and we define $\Gamma_n^{F,\ell}$ on $[0,1]^d$, see Theorems 4.9 and 4.11 for the multidimensional case. The case $g = \mathrm{Id}$ appears already in [84], see Corollary 4.13 for the graphs indexed version. We have the following convergence of finite-dimensional distributions (or equivalently of the process since $\mathcal{F}$ is countable).

Corollary (Corollary 4.13)**.** *We have the following convergence of finite-dimensional distributions:*
$$\left(\sqrt{n}\left(t_{\mathrm{inj}}(F, G_n) - t(F, W)\right) : F \in \mathcal{F}\right) \xrightarrow[n\to\infty]{(fdd)} \Theta_{\mathrm{inj}},$$
where $\Theta_{\mathrm{inj}} = (\Theta_{\mathrm{inj}}(F) : F \in \mathcal{F})$ is a centered Gaussian process with covariance function K_{inj} given, for $F, F' \in \mathcal{F}$, by:
$$K_{\mathrm{inj}}(F, F') = \sum_{q \in V(F)} \sum_{q' \in V(F')} t\left((F \bowtie F')(q, q'), W\right) - v(F)v(F')\, t(F, W)t(F', W).$$

As a consequence, we get the central limit theorem for homomorphism densities from quantum graphs, see (4.52) and for induced homomorphism densities, see Corollary 4.15. In the Erdős-Rényi case, the one-dimensional limit distribution of induced homomorphism densities is not necessarily normal: its behaviour depends on the number of edges, two-stars and triangles in the graph F, see [159] and [118].

Notice that because $g = \mathbf{1}_{[0,D(y)]}$ is not of class $\mathcal{C}^2([0,1])$, we can not apply Theorem 4.11 (with $F = K_2$ and $k = 1$) directly to get the convergence in distribution of $\sqrt{n}(\Pi_n(y) - y)$ towards $\chi(y)$ given in Theorem 4.23. Nevertheless, the asymptotic variance can be formally obtained by computing $\sigma^{K_2,\bullet}(g)$ given in Theorem 4.11 with $g = \mathbf{1}_{[0,D(y)]}$ and $g'(z)dz = (D'(y))^{-1}\delta_{D(y)}(dz)$, with $\delta_{D(y)}(dz)$ the Dirac mass at $D(y)$. However, the proofs of Theorems 4.11 and 4.23 require different approaches.

Similarly to Theorem 4.23 and in the spirit of Theorem 4.11, it could be interesting to consider the convergence of CDF for triangles or more generally for simple finite graphs F, $V(F) = [p]$, and $\ell \in \mathcal{S}_{p,k}$:

$$\left(\frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \mathbf{1}_{\left\{ t_{\mathrm{inj}}(F^\ell, G_n^\alpha) \le t_x(F^\ell, W) \right\}} : x \in (0,1)^k \right),$$

where $t_x(F^\ell, W) = \mathbb{E}\left[t_{\mathrm{inj}}\left(F^\ell, G_n^{[k]}\right) \middle| (X_1, \dots, X_k) = x \right]$, see (4.31) and the second equality in (4.37).

4.1.4 Organization of the paper

We recall the definitions of graph homomorphisms, graphons, W-random graphs in Section 4.2. We present our main result about the almost sure convergence for the random measure $\Gamma_n^{F,\ell}$ associated to homomorphism densities of sampling partially labeled graphs from a graphon in Section 4.3.2, see Theorem 4.9. The proof is given in Section 4.5 after a preliminary result given in Section 4.4. The associated fluctuations are stated in Theorem 4.11 and proved in Section 4.6. Section 4.7 is devoted to the asymptotics for the empirical CDF of degrees Π_n, see Theorem 4.23 for the fluctuations corresponding to the almost sure convergence. After some ancillary results given in Section 4.8, we prove Theorem 4.23 in Section 4.9. We add an index of notation at the end of the paper for the reader's convenience. We postpone to the appendices some technical results on precise uniform asymptotics for the CDF of binomial distributions, see Section 4.10, and a proof of Proposition 4.28 on approximation for the CDF of multivariate binomial distributions.

4.2 Definitions

4.2.1 First notation

We denote by $|B|$ the cardinality of the set B. For $n \in \mathbb{N}^*$, we set $[n] = \{1, \dots, n\}$. Let $\mathcal{A}$ be a non-empty set of characters, called the alphabet. A sequence $\beta = \beta_1 \dots \beta_k$, with $\beta_i \in \mathcal{A}$ for all $1 \le i \le k$, is called a $\mathcal{A}$-word (or string) of length $|\beta| = k \in \mathbb{N}^*$. The word β is also identified with the vector $(\beta_1, \dots, \beta_k)$, and for $q \in \mathcal{A}$, we write $q \in \beta$ if q belongs to $\{\beta_1, \dots, \beta_k\}$. The concatenation of two $\mathcal{A}$-words α and β is denoted by $\alpha\,\beta$.

We now define several other operations on words. Let β be a $\mathcal{A}$-word of length $p \in \mathbb{N}^*$ and $k \in [p]$. For α a $[p]$-word of length k, we consider the $\mathcal{A}$-word β_α, defined by

$$\beta_\alpha = \beta_{\alpha_1} \dots \beta_{\alpha_k}. \tag{4.4}$$

The word $\beta_{[k]} = \beta_1 \dots \beta_k$ corresponds to the first k terms of β, where by convention, $[k]$ denotes the $\mathbb{N}^*$-word $1 \dots k$. We define, for $i, j \in [p]$, the transposition word $\tau_{ij}(\beta)$ of β, obtained by exchanging the place of the ith character with the jth character in the word β: for $u \in [p]$,

$$\tau_{ij}(\beta)_u = \begin{cases} \beta_u & \text{if } u \notin \{i, j\}, \\ \beta_i & \text{if } u = j, \\ \beta_j & \text{if } u = i. \end{cases} \tag{4.5}$$

Finally, for $q \in \mathcal{A}$ and $i \in [p]$, we define the new $\mathcal{A}$-word $R_i(\beta, q)$, derived from β by substituting its ith character with q: for $u \in [p]$,

$$R_i(\beta, q)_u = \begin{cases} \beta_u & \text{if } u \ne i, \\ q & \text{if } u = i. \end{cases} \tag{4.6}$$

Let $n \in \mathbb{N}^*$ and $p \in [n]$. We define the set $\mathcal{S}_{n,p}$ of all $[n]$-words of length p such that all characters are distinct:

$$\mathcal{S}_{n,p} = \{\beta = \beta_1 \dots \beta_p : \beta_i \in [n] \text{ for all } i \in [p] \text{ and } \beta_1, \dots, \beta_p \text{ are all distinct}\}. \tag{4.7}$$

Notice that $|\mathcal{S}_{n,p}| = A_n^p = n!/(n-p)!$, and that $\mathcal{S}_{n,1} = [n]$. Moreover, for $n \in \mathbb{N}^*$, $\mathcal{S}_{n,n}$ is simply the set of all permutations of $[n]$ which will be also denoted by $\mathcal{S}_n$. With this notation, for $n \in \mathbb{N}^*$, we define the set $\mathcal{M}_n$ of all $[n]$-words with all characters distinct:

$$\mathcal{M}_n = \bigcup_{p \in [n]} \mathcal{S}_{n,p}. \tag{4.8}$$

Let $n \geq p \geq k \geq 1$ and $\ell \in \mathcal{S}_{p,k}$. For $\alpha \in \mathcal{S}_{n,k}$, we define the set $\mathcal{S}_{n,p}^{\ell,\alpha}$ of all $[n]$-words of length p such that all characters are distinct and for all $i \in [k]$, the ℓ_i-th character is equal to α_i:

$$\mathcal{S}_{n,p}^{\ell,\alpha} = \{\beta \in \mathcal{S}_{n,p} : \beta_\ell = \alpha\} \cdot \tag{4.9}$$

We have $|\mathcal{S}_{n,p}^{\ell,\alpha}| = A_{n-k}^{p-k}$. As $A_n^p = A_n^k A_{n-k}^{p-k}$, that is $|\mathcal{S}_{n,p}| = |\mathcal{S}_{n,k}|\,|\mathcal{S}_{n,p}^{\ell,\alpha}|$ for any $\alpha \in \mathcal{S}_{n,k}$, we get that for all real-valued functions f defined on $\mathcal{S}_{n,k}$:

$$\frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} f(\beta_\ell) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} f(\alpha) \cdot \tag{4.10}$$

Let $d \in \mathbb{N}^*$. For $x, y \in \mathbb{R}^d$, we denote by $\langle x, y \rangle$ the usual scalar product on $\mathbb{R}^d$ and $|x| = \sqrt{\langle x, x \rangle}$ the Euclidean norm in $\mathbb{R}^d$.

We use the convention $\prod_\emptyset = 1$.

4.2.2 Graph homomorphisms

A simple finite graph G is an ordered pair $(V(G), E(G))$ of a set $V(G)$ of $v(G)$ vertices, and a subset $E(G)$ of the collection of $\binom{v(G)}{2}$ unordered pairs of vertices. We usually shall identify $V(G)$ with $[v(G)]$. The elements of $E(G)$ are called edges and we denote by $e(G) = |E(G)|$ the number of edges in the graph G. Recall a graph G is simple when it has no self-loops, and no multiple edges between any pair of vertices. Let $\mathcal{F}$ be the set of all simple finite graphs.

Let $F, G \in \mathcal{F}$ be two simple finite graphs and set $p = v(F)$ and $n = v(G)$. A homomorphism f from F to G is an adjacency-preserving map from $V(F) = [p]$ to $V(G) = [n]$ i.e. a map from $V(F)$ to $V(G)$ such that if $\{i, j\} \in E(F)$ then $\{\mathrm{f}(i), \mathrm{f}(j)\} \in E(G)$. Let $\mathrm{Hom}(F, G)$ denote the set of homomorphisms from F to G. The homomorphism density from F to G is the normalized quantity:

$$t(F, G) = \frac{|\mathrm{Hom(F,G)}|}{n^p}. \tag{4.11}$$

It is the probability that a uniform random map from $V(F)$ to $V(G)$ is a homomorphism. We have a similar definition when f is restricted to being injective. Let $\mathrm{Inj}(F, G)$ denote the set of injective homomorphisms of F into G and define its density as:

$$t_{\mathrm{inj}}(F, G) = \frac{|\mathrm{Inj}(F, G)|}{A_n^p}. \tag{4.12}$$

For $\beta \in \mathcal{S}_{n,p}$, we set, with $V(F) = [p]$ and $V(G) = [n]$:

$$Y^\beta(F, G) = \prod_{\{i,j\} \in E(F)} \mathbf{1}_{\{\{\beta_i, \beta_j\} \in E(G)\}}. \tag{4.13}$$

When there is no risk of confusion, we shall write Y^β for $Y^\beta(F,G)$, and thus we have:

$$t_{\text{inj}}(F,G) = \frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} Y^\beta. \tag{4.14}$$

We recall from Lovász [138], Section 5.2.3, that:

$$|t_{\text{inj}}(F,G) - t(F,G)| \leq \frac{1}{n}\binom{p}{2}. \tag{4.15}$$

In the same way, we can define homomorphism densities from partially labeled graphs. See Figure 4.1 for an injective homomorphism of partially labeled graphs. Assume $p > k \geq 1$. Let $\ell \in \mathcal{S}_{p,k}$ and $\alpha \in \mathcal{S}_{n,k}$. We define $\text{Inj}(F^\ell, G^\alpha)$ the set of injective homomorphisms f from F into G such that $\text{f}(\ell_i) = \alpha_i$ for all $i \in [k]$, and its density:

$$t_{\text{inj}}(F^\ell, G^\alpha) = \frac{|\text{Inj}(F^\ell, G^\alpha)|}{A_{n-k}^{p-k}} = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} Y^\beta. \tag{4.16}$$

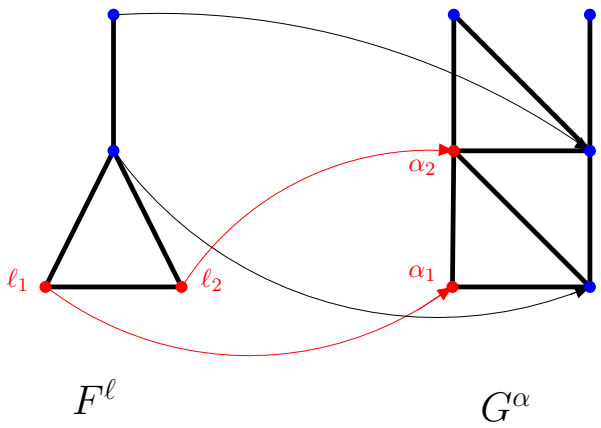


Figure 4.1 – Example of an injective homomorphism from partially labeled graphs.

Denote $\mathcal{R}_\ell(F)$ the labeled sub-graph of F with vertices $\{\ell_1, \ldots, \ell_k\}$ and edges:

$$E(\mathcal{R}_\ell(F)) = \{\{i,j\} \in E(F) : i,j \in \ell\}. \tag{4.17}$$

For $\alpha \in \mathcal{S}_{n,k}$, we set:

$$\hat{Y}^\alpha(F^\ell, G^\alpha) = Y^\alpha(\mathcal{R}_\ell(F), G) = \prod_{\{i,j\} \in E(\mathcal{R}_\ell(F))} \mathbf{1}_{\{\{\alpha_i,\alpha_j\} \in E(G)\}}, \tag{4.18}$$

For $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we set $Y^\beta(F^\ell, G^\alpha) = \hat{Y}^\alpha(F^\ell, G^\alpha)\, \tilde{Y}^\beta(F^\ell, G^\alpha)$ with:

$$\tilde{Y}^\beta(F^\ell, G^\alpha) = \prod_{\{i,j\} \in E(F) \setminus E(\mathcal{R}_\ell(F))} \mathbf{1}_{\{\{\beta_i,\beta_j\} \in E(G)\}}. \tag{4.19}$$

Notice that $Y^\beta(F^\ell, G^\alpha)$ is equal to Y^β defined in (4.13) for $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$. When there is no risk of confusion, we shall write $\hat{Y}^\alpha, \tilde{Y}^\beta$ and Y^β for $\hat{Y}^\alpha(F^\ell, G^\alpha), \tilde{Y}^\beta(F^\ell, G^\alpha)$ and $Y^\beta(F^\ell, G^\alpha)$. Remark that $\hat{Y}^\alpha$ is either 0 or 1. By construction, we have:

$$t_{\text{inj}}(F^\ell, G^\alpha) = \hat{Y}^\alpha\, \tilde{t}_{\text{inj}}(F^\ell, G^\alpha) \quad \text{with} \quad \tilde{t}_{\text{inj}}(F^\ell, G^\alpha) = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} \tilde{Y}^\beta. \tag{4.20}$$

Summing (4.16) over $\alpha \in \mathcal{S}_{n,k}$, we get using (4.10) and (4.12) that:

$$\frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\mathrm{inj}}(F^\ell, G^\alpha) = t_{\mathrm{inj}}(F, G). \tag{4.21}$$

We can generalize this formula as follows. Let $n \geq p > k > k' \geq 1$, $\ell \in \mathcal{S}_{p,k}$, $\gamma \in \mathcal{S}_{k,k'}$ and $\alpha' \in \mathcal{S}_{n,k'}$. We easily get:

$$\frac{1}{|\mathcal{S}_{n,k}^{\gamma,\alpha'}|} \sum_{\alpha \in \mathcal{S}_{n,k}^{\gamma,\alpha'}} t_{\mathrm{inj}}(F^\ell, G^\alpha) = t_{\mathrm{inj}}(F^{\ell_\gamma}, G^{\alpha'}). \tag{4.22}$$

Remark 4.1. Let K_2 (resp. $K_2^\bullet$) denote the complete graph with two vertices (resp. one of them being labeled). Let $G \in \mathcal{F}$ with n vertices. We define the degree sequence $(D_i(G) : i \in [n])$ of the graph G by, for $i \in [n]$:

$$D_i(G) = t_{\mathrm{inj}}\left(K_2^\bullet, G^i\right) = \frac{1}{n-1} \sum_{j \in [n] \setminus \{i\}} \mathbf{1}_{\{\{i,j\} \in E(G)\}}. \tag{4.23}$$

Remark 4.2. Let $F \in \mathcal{F}$ be a simple finite graph with $V(F) = [p]$. Let $\ell \in \mathcal{S}_{p,k}$ for some $k \in [p]$. Assume F_0 is obtained from F by adding p' isolated vertices numbered from $p+1$ to $p+p'$, and label ℓ' of those isolated vertices so that ℓ' is a $\{p+1, \ldots, p+p'\}$-word of length say $k' = |\ell'| \leq p'$ and $\ell\ell' \in \mathcal{S}_{p+p',k+k'}$. By convention $k' = 0$ means none of the added isolated vertices is labeled. Assume $n \geq p + p'$ and let $\alpha \in \mathcal{S}_{n,k}$ and α' be a $[n]$-word such that $\alpha\alpha' \in \mathcal{S}_{n,k+k'}$. Then, it is elementary to check that:

$$t_{\mathrm{inj}}(F_0^{\ell\ell'}, G^{\alpha\alpha'}) = t_{\mathrm{inj}}(F^\ell, G^\alpha).$$

as well as, with δ_x the Dirac mass at x:

$$\frac{1}{|\mathcal{S}_{n,k+k'}|} \sum_{\alpha\alpha' \in \mathcal{S}_{n,k+k'}} \delta_{t_{\mathrm{inj}}(F_0^{\ell\ell'}, G^{\alpha\alpha'})} = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \delta_{t_{\mathrm{inj}}(F^\ell, G^\alpha)}.$$

In conclusion adding isolated vertices (labeled or non labeled) does not change the homomorphism densities.

Finally, we recall an induced homomorphism from F to G is an injective homomorphism which preserves non-adjacency, that is: an injective maps f from $V(F)$ to $V(G)$ is an induced homomorphism if $\{i,j\} \in E(F)$ if and only if $\{\mathrm{f}(i), \mathrm{f}(j)\} \in E(G)$. See Figure 4.2 for an injective homomorphism which is not an induced homomorphism. Let $\mathrm{Ind}(F,G)$ denote the set of induced homomorphisms; we denote its density by:

$$t_{\mathrm{ind}}(F, G) = \frac{|\mathrm{Ind}(F, G)|}{A_n^p}. \tag{4.24}$$

We recall results from [138], see Section 5.2.3., which gives relations between injective and induced homomorphism densities.

Proposition 4.3. *For $F, G \in \mathcal{F}$, two simple finite graphs, we have:*

$$t_{\mathrm{inj}}(F, G) = \sum_{F' \geq F} t_{\mathrm{ind}}(F', G) \quad \text{and} \quad t_{\mathrm{ind}}(F, G) = \sum_{F' \geq F} (-1)^{e(F')-e(F)} t_{\mathrm{inj}}(F', G), \tag{4.25}$$

where $F' \geq F$ means that $V(F) = V(F')$ and $E(F) \subset E(F')$, that is F' ranges over all simple graphs obtained from F by adding edges.

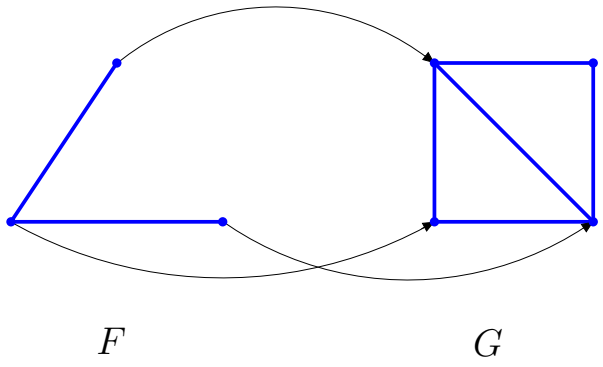


Figure 4.2 – An example of an injective homomorphism but not an induced homomorphism.

4.2.3 Graphons

A graphon is a symmetric, measurable function $W : [0,1]^2 \to [0,1]$. Denote the space of all graphons by $\mathcal{W}$. Homomorphism densities from graphs can be extended to graphons. For every simple finite graph F and every graphon $W \in \mathcal{W}$, we define

$$t(F,W) = t_{\text{inj}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i,x_j) \prod_{k\in V(F)} dx_k \tag{4.26}$$

and

$$t_{\text{ind}}(F,W) = \int_{[0,1]^{V(F)}} \prod_{\{i,j\}\in E(F)} W(x_i,x_j) \prod_{\{i,j\}\notin E(F)} (1-W(x_i,x_j)) \prod_{k\in V(F)} dx_k. \tag{4.27}$$

A sequence of simple finite graphs $(H_n : n \in \mathbb{N}^*)$ is called convergent if the sequence $(t(F,H_n) : n \in \mathbb{N}^*)$ has a limit for every simple finite graph F. Lovász and Szegedy [139] proved that the limit of a convergent sequence of graphs can be represented as a graphon, up to a measure preserving bijection. In particular, a sequence of graphs $(G_n : n \in \mathbb{N}^*)$ is said to converge to a graphon W if for every simple finite graph F, we have

$$\lim_{n\to\infty} t(F,H_n) = t(F,W).$$

As an extension, we can define homomorphism densities from a k-labeled simple finite graph F to a graphon W which are defined by not integrating the variables corresponding to labeled vertices. Let $F \in \mathcal{F}$ be a simple finite graph, set $p = v(F)$ and identify $V(F)$ with $[p]$. Let $p \geq k \geq 1$ and $\ell \in \mathcal{S}_{n,k}$. Recall $E(\mathcal{R}_\ell(F))$ defined in (4.17). We set for $y = (y_1,\ldots,y_p) \in [0,1]^p$:

$$\tilde{Z}(y) = \prod_{\{i,j\}\in E(F)\setminus E(\mathcal{R}_\ell(F))} W(y_i,y_j) \tag{4.28}$$

and for $x = (x_1,\ldots,x_k) \in [0,1]^k$ we consider the average of $\tilde{Z}(y)$ over y restricted to $y_\ell = x$:

$$\tilde{t}_x(F^\ell,W) = \int_{[0,1]^p} \tilde{Z}(y) \prod_{m\in[p]\setminus\ell} dy_m \prod_{m'\in[k]} \delta_{x_{m'}}(dy_{\ell_{m'}}), \tag{4.29}$$

as well as the analogue of $\hat{Y}^\alpha$ for the graphon:

$$\hat{t}_x(F^\ell,W) = \prod_{\{\ell_i,\ell_j\}\in E(\mathcal{R}_\ell(F))} W(x_i,x_j) \quad \text{and} \quad \hat{t}(F^\ell,W) = \int_{[0,1]^k} \hat{t}_x(F^\ell,W)\,dx = t(\mathcal{R}_\ell(F),W). \tag{4.30}$$

Similarly to (4.20), we set for $\ell \in \mathcal{S}_{n,k}$ and $x \in [0,1]^k$:

$$t_x(F^\ell, W) = \hat{t}_x(F^\ell, W)\, \tilde{t}_x(F^\ell, W). \tag{4.31}$$

Let β and β' be $[k]$-words such that $\beta\beta' \in \mathcal{S}_k$, with $k' = |\beta'|$ and $1 \le k' < k$. We easily get:

$$\int_{[0,1]^{k'}} t_x(F^\ell, W)\, dx_{\beta'} = t_{x_\beta}(F^{\ell_\beta}, W). \tag{4.32}$$

The result also holds for $k' = k$ with the convention $t_{x_\beta}(F^{\ell_\beta}, W) = t(F, W)$ when $\beta = \emptyset$.

Remark 4.4. The Erdős-Rényi case corresponds to $W \equiv \mathfrak{p}$ with $0 < \mathfrak{p} < 1$, and in this case we have $t(F, W) = t_x(F^\ell, W) = \mathfrak{p}^{e(F)}$ for all $x \in [0,1]^k$.

Remark 4.5. The normalized degree function D of the graphon W is defined by, for all $x \in [0,1]$:

$$D(x) = \int_0^1 W(x,y)dy. \tag{4.33}$$

We have for $W \in \mathcal{W}$ and $x \in [0,1]$:

$$t_x(K_2^\bullet, W) = \int_0^1 W(x,y)dy = D(x) \quad \text{and} \quad t(K_2, W) = \int_0^1 D(x)dx.$$

4.2.4 W-random graphs

To complete the identification of graphons as the limit object of convergent sequences, it has been proved by Lovász and Szegedy [139] that we can always find a sequence of graphs, given by a sampling method, whose limit is a given graphon function.

Let $W \in \mathcal{W}$. We can generate a W-random graph G_n with vertex set $[n]$ from the given graphon W, by first taking an independent sequence $X = (X_i : i \in \mathbb{N}^*)$ with uniform distribution on $[0,1]$, and then, given this sequence, letting $\{i,j\}$ with $i,j \in [n]$ be an edge in G_n with probability $W(X_i, X_j)$. When we need to stress the dependence on W, we shall write $G_n(W)$ for G_n. For a given sequence X, this is done independently for all pairs $(i,j) \in [n]^2$ with $i < j$.

The random graphs $G_n(W)$ thus generalize the Erdős-Rényi random graphs $G_n(\mathfrak{p})$ obtained by taking $W \equiv \mathfrak{p}$ with $0 < \mathfrak{p} < 1$ constant. (We recall that the Erdős-Rényi random graph $G_n(\mathfrak{p})$ is a random graph defined on the finite set $[n]$ of vertices whose edges occur independently with the same probability $\mathfrak{p}$, $0 < \mathfrak{p} < 1$.) Moreover, $(G_n : n \in \mathbb{N}^*)$ converges a.s. towards the graphon W, see for instance [138], Proposition 11.32.

Remark 4.6. We provide elementary computations which motivate the introduction in the previous section of $\hat{t}_x(F^\ell, W)$ and $\tilde{t}_x(F^\ell, W)$. Recall that $X_\gamma = (X_{\gamma_1}, \ldots, X_{\gamma_r})$ with γ a $\mathbb{N}^*$-word of length $|\gamma| = r$. Let $n \ge p \ge 1$ and $F \in \mathcal{F}$ with $V(F) = [p]$ and $\ell \in \mathcal{S}_{p,k}$. We set for $x = (x_1, \ldots, x_p) \in [0,1]^p$:

$$Z(x) = \prod_{\{i,j\} \in E(F)} W(x_i, x_j).$$

Let $\alpha \in \mathcal{S}_{p,k}$ and $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$. By construction, we have:

$$Z(X_\beta) = \mathbb{E}\left[Y^\beta(F, G_n) \,|\, X\right] = \mathbb{E}\left[Y^\beta(F^\ell, G_n^\alpha) \,|\, X\right].$$

By definition of $\hat{t}_x(F^\ell, W)$ and $\tilde{t}_x(F^\ell, W)$, we get:

$$\hat{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha) \,|\, X\right] = \mathbb{E}\left[Y^\alpha(\mathcal{R}_\ell(F), G_n^\alpha) \,|\, X\right] \tag{4.34}$$

$$\tilde{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\tilde{Z}(X_\beta) \,|\, X_\alpha\right] = \mathbb{E}\left[\tilde{Y}^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \tag{4.35}$$

$$t_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[Z(X_\beta) \,|\, X_\alpha\right] = \mathbb{E}\left[Y^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right]. \tag{4.36}$$

By summing (4.35) and (4.36) over $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we get, using (4.20) and (4.17), that

$$\tilde{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \quad \text{and} \quad t_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[t_{\mathrm{inj}}(F^\ell, G_n^\alpha) \,|\, X_\alpha\right]. \tag{4.37}$$

Taking the expectation in the second equality of (4.37), we deduce that:

$$t(F, W) = \int_{[0,1]^k} t_x(F^\ell, W)\, dx = \mathbb{E}\left[t_{X_\alpha}(F^\ell, W)\right] = \mathbb{E}\left[t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right].$$

Thanks to (4.21), we recover that

$$t(F, W) = \mathbb{E}\left[t_{\mathrm{inj}}(F, G_n)\right],$$

see also [138], Proposition 11.32 or [143] Proposition A.1. We also have:

$$t(F, W) = \mathbb{E}\left[Z(X_\beta)\right] = \mathbb{E}\left[Y^\beta(F, W)\right].$$

By definition of $\hat{t}(F^\ell, W)$, we get:

$$\hat{t}(F^\ell, W) = \mathbb{E}\left[\hat{t}_{X_\alpha}(F^\ell, W)\right] = \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha)\right] = \mathbb{E}\left[Y^\alpha(\mathcal{R}_\ell(F), G_n)\right] = t(\mathcal{R}_\ell(F), W).$$

Since $\hat{Y}^\alpha(F^\ell, G_n^\alpha)$ and $\tilde{Y}^\beta(F^\ell, G_n^\alpha)$ are, conditionally on X or X_β or X_α, independent, we deduce that:

$$\begin{aligned} t_{X_\alpha}(F^\ell, W) &= \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha)\tilde{Y}^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \\ &= \mathbb{E}\left[\hat{Y}^\alpha(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \mathbb{E}\left[\tilde{Y}^\beta(F^\ell, G_n^\alpha) \,|\, X_\alpha\right] \\ &= \hat{t}_{X_\alpha}(F^\ell, W)\ \tilde{t}_{X_\alpha}(F^\ell, W). \end{aligned}$$

This latter equality gives another interpretation of (4.31).

4.3 Asymptotics for homomorphism densities of sampling partially labeled graphs from a graphon

4.3.1 Random measures associated to a graphon

Let $d \geq 1$ and $I = [0,1]^d$. We denote by $\mathcal{B}(I)$ (resp. $\mathcal{B}^+(I)$) the set of all real-valued (resp. non negative) measurable functions defined on I. We denote by $\mathcal{C}(I)$ (resp. $\mathcal{C}_b(I)$) the set of real-valued (resp. bounded) continuous functions defined on I. For $f \in \mathcal{B}(I)$ we denote by $\|f\|_\infty$ the supremum norm of f on I. We denote by $\mathcal{C}^k(I)$ the set of real-valued functions f defined on I with continuous k-th derivative. For $f \in \mathcal{C}^1(I)$, its derivative is denoted by $\nabla f = (\nabla_1 f, \ldots, \nabla_d f)$ and we set $\|\nabla f\|_\infty = \sum_{i=1}^d \|\nabla_i f\|_\infty$.

Let $F = (F_m : 1 \leq m \leq d) \in \mathcal{F}^d$ be a finite sequence of simple finite graphs. Using Remark 4.2, if necessary, we can complete the graphs F_m with isolated vertices such that for all $m \in [d]$, we have $v(F_m) = p$ for some $p \in \mathbb{N}^*$ and consider that $V(F_m) = [p]$. We shall write $p = v(F)$. Let $\ell \in \mathcal{M}_p$ (where $\mathcal{M}_p$ is the set of all $[p]$-words with all characters distinct, given by (4.8)) and set $k = |\ell|$. We denote:

$$t_{\mathrm{inj}}(F^\ell, G_n^\alpha) = \left(t_{\mathrm{inj}}(F_m^\ell, G_n^\alpha) : m \in [d]\right) \in [0,1]^d,$$

and similarly for $\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)$. Let W be a graphon and $x \in [0,1]^k$. Similarly, we define $t_x(F^\ell, W)$, and $\tilde{t}_x(F^\ell, W)$, so for example:

$$t_x(F^\ell, W) = \left(t_x(F_m^\ell, W) : m \in [d]\right) \in [0,1]^d.$$

Notice that relabeling F_m if necessary, we get all the possible combinations of density of labeled injective homomorphism of F_m into G for all $m \in [d]$ (we could even take $\ell = [k]$).

Recall $F_m^{[\ell]}$ is the labeled sub-graph of F_m with vertices $\{\ell_1, \ldots, \ell_k\}$ and set of edges $E(F_m^{[\ell]}) = \{\{i,j\} \in E(F_m) : i,j \in \ell\}$ see (4.17). For simplicity, we shall assume the following condition which states that $F_m^{[\ell]}$ does not depend on m:

$$\text{For } m, m' \in [d],\ i, i' \in \ell, \text{ we have: } \{i,i'\} \in E(F_m) \Longleftrightarrow \{i,i'\} \in E(F_{m'}). \tag{4.38}$$

This condition can be removed when stating the main results from Section 4.3.2 at the cost of very involved notation. Therefore, we shall leave this extension to the very interested reader.

Let $G_n = G_n(W)$ be the associated W-random graphs with n vertices constructed from W and the sequence $X = (X_i : i \in \mathbb{N}^*)$ of independent uniform random variables on $[0,1]$. Under Condition (4.38), for $\alpha \in \mathcal{S}_{n,k}$ and $x \in [0,1]^k$, we have that $\hat{Y}^\alpha(F_m^\ell, G_n^\alpha)$ and $\hat{t}_x(F_m^\ell, W)$ do not depend on $m \in [d]$. We set $\hat{Y}^\alpha(F^\ell, G_n^\alpha)$ and $\hat{t}_x(F^\ell, W)$ for the common values. When there is no confusion, we write $\hat{Y}^\alpha$ for $\hat{Y}^\alpha(F^\ell, G_n^\alpha)$. In particular, we deduce from (4.20) that:

$$t_{\mathrm{inj}}(F^\ell, G_n^\alpha) = \hat{Y}^\alpha\ \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) \quad \text{with} \quad \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) = \left(\tilde{t}_{\mathrm{inj}}(F_m^\ell, G_n^\alpha) : m \in [d]\right). \tag{4.39}$$

Remark 4.7. If $|\ell| = k = 1$, then Condition (4.38) is automatically satisfied and we have by convention that $\hat{Y}^\alpha = \hat{t}_x(F^\ell, W) = 1$ for $\alpha \in \mathcal{S}_{n,k}$ and $x \in [0,1]^k$. If $d = 1$, then, Condition (4.38) is also automatically satisfied.

We define the random probability measure $\Gamma_n^{F,\ell}$ on $([0,1]^d, \mathcal{B}([0,1]^d))$ by, for $g \in \mathcal{B}^+([0,1]^d)$:

$$\begin{aligned}\Gamma_n^{F,\ell}(g) &= \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} g\left(t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) \\ &= \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha g\left(\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) + (1 - \hat{Y}^\alpha) g(0),\end{aligned} \tag{4.40}$$

where we used (4.39) and the fact that $\hat{Y}^\alpha$ takes values in $\{0,1\}$ for the second equality. For $k \in \mathbb{N}^*$ and α an $\mathbb{N}^*$-word of length k, we recall the notation $X_\alpha = (X_{\alpha_1}, \ldots, X_{\alpha_k})$ and $X_{[k]} = (X_1, \ldots, X_k)$. Recall (4.29) and (4.30). We define the auxiliary random probability measure $\hat{\Gamma}_n^{F,\ell}$ on $[0,1]^d$ by, for $g \in \mathcal{B}^+([0,1]^d)$:

$$\hat{\Gamma}_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{t}_{X_\alpha}(F^\ell, W)\, g\left(\tilde{t}_{X_\alpha}(F^\ell, W)\right) + \left(1 - \hat{t}_{X_\alpha}(F^\ell, W)\right) g(0). \tag{4.41}$$

and the deterministic probability measure $\Gamma^{F,\ell}$, by, for all $g \in \mathcal{B}^+([0,1]^d)$:

$$\begin{aligned}\Gamma^{F,\ell}(g) &= \mathbb{E}\left[\hat{\Gamma}_n^{F,\ell}(g)\right] \\ &= \int_{[0,1]^k} \hat{t}_x(F^\ell, W)\, g\left(\tilde{t}_x(F^\ell, W)\right)\, dx + \left(1 - \hat{t}(F^\ell, W)\right) g(0).\end{aligned} \tag{4.42}$$

Remark 4.8.

(i) If $d = 1$ and $g = \mathrm{Id}$, then we have thanks to (4.21) that:

$$\Gamma_n^{F,\ell}(\mathrm{Id}) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} t_{\mathrm{inj}}(F^\ell, G_n^\alpha) = t_{\mathrm{inj}}(F, G_n)$$

and, thanks to (4.31) and (4.32):

$$\Gamma^{F,\ell}(\mathrm{Id}) = \int_{[0,1]^k} t_x(F^\ell, W)dx = t(F, W).$$

Notice that $\Gamma_n^{F,\ell}(\mathrm{Id})$ and $\Gamma^{F,\ell}(\mathrm{Id})$ do not depend on ℓ.

(ii) If $|\ell| = 1$, then according to Remark 4.7, we get:

$$\Gamma^{F,\ell}(g) = \int_{[0,1]} g\left(t_x(F^\ell, W)\right)\, dx.$$

4.3.2 Invariance principle and its fluctuations

We first state the invariance principle for the random probability measure $\Gamma_n^{F,\ell}$. The proof of the next theorem is given in Section 4.5.

Theorem 4.9. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $V(F) = [p]$, $\ell \in \mathcal{M}_p$. Assume that Condition (4.38) holds. Then, the sequence of random probability measures on $[0,1]^d$, $\left(\Gamma_n^{F,\ell} : n \in \mathbb{N}^*\right)$ converges a.s. for the weak topology towards $\Gamma^{F,\ell}$.*

The convergence of $\left(\Gamma_n^{F,\ell}(\mathrm{Id}) : n \in \mathbb{N}^*\right)$, with $d = 1$, can also be found in [138], see Proposition 11.32.

Remark 4.10. By Portmanteau Theorem, we have that a.s. for all bounded measurable function g on $[0,1]^d$ such that $\Gamma^{F,\ell}(\mathcal{D}_g) = 0$ where $\mathcal{D}_g$ is the set of discontinuity points of g, $\lim_{n\to\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.

For simplicity, consider the case $d = 1$ and $W \equiv \mathfrak{p}$ with $0 < \mathfrak{p} < 1$. Let $\hat{e}(F)$ denote the cardinality of $E(\mathcal{R}_\ell(F))$. Because $\Gamma^{F,\ell} = \mathfrak{p}^{\hat{e}(F)}\delta_{\mathfrak{p}^{e(F)-\hat{e}(F)}} + (1 - \mathfrak{p}^{\hat{e}(F)})\delta_0$, with $k = |\ell|$, then if g is continuous at $\mathfrak{p}^{e(F)-\hat{e}(F)}$ and at 0, we get that a.s. $\lim_{n\to\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.

The next theorem, whose proof is given in Section 4.6, gives the fluctuations corresponding to the invariance principle of Theorem 4.9. Notice the speed of convergence in the invariance principle is of order $\sqrt{n}$.

For $\mu \in \mathbb{R}$ and $\sigma \geq 0$, we denote by $\mathcal{N}(\mu, \sigma^2)$ the Gaussian distribution with mean μ and variance σ^2.

Theorem 4.11. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $V(F) = [p]$, $\ell \in \mathcal{M}_p$, with $k = |\ell|$. Assume that Condition (4.38) holds. Then, for all $g \in \mathcal{C}^2([0,1]^d)$, we have the following convergence in distribution:*

$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$

with $\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(\mathcal{U}_g^{F,\ell}\right)$ and

$$\mathcal{U}_g^{F,\ell} = \sum_{i=1}^{k} \int_{[0,1]^k} \hat{t}_{R_i(x,U)}(F^\ell, W)\, \left(g\big(\tilde{t}_{R_i(x,U)}(F^\ell, W)\big) - g(0)\right) dx$$
$$+ \sum_{q\in[p]\setminus\ell} \int_{[0,1]^k} dx\, \langle t_{xU}(F^{\ell q}, W), \nabla g\left(\tilde{t}_x(F^\ell, W)\right)\rangle, \quad (4.43)$$

where U is a uniform random variable on $[0,1]$, and $[p]\setminus\ell = \{1,\ldots,p\}\setminus\{\ell_1,\ldots,\ell_k\}$.

Remark 4.12. Let U be a uniform random variable on $[0,1]$.

(i) Assume that $|\ell| = k = 1$. Using Remark 4.7, we get for $\ell \in [p]$:

$$\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(g\left(t_U(F^\ell, W)\right) + \sum_{q \in [p]\setminus \ell} \int_{[0,1]} \langle t_{xU}(F^{\ell q}, W), \nabla g\big(t_x(F^\ell, W)\big)\rangle \, dx \right). \tag{4.44}$$

(ii) Let $F \in \mathcal{F}^d$ with $p = v(F)$. Take $\ell = 1$ with $k = |\ell| = 1$. Let $a \in \mathbb{R}^d$ and consider $g(x) = \langle a, x\rangle$ for $x \in \mathbb{R}^d$. We deduce from (4.44) and (4.32) that:

$$\sigma^{F,\ell}(g)^2 = \mathrm{Var}\left(\langle a, \sum_{q=1}^{p} t_U\left(F^q, W\right)\rangle \right). \tag{4.45}$$

(iii) In the case $d = 1$, $F \in \mathcal{F}$, and $g = \mathrm{Id}$, the central limit theorem appears already in [84]. In this case, we have $\Gamma_n^{F,\ell}(\mathrm{Id}) = t_{\mathrm{inj}}(F, G_n)$, $\Gamma^{F,\ell}(\mathrm{Id}) = t(F, W)$ and, thanks to (4.45) (with $d = 1$ and $a = 1$):

$$\sigma^{F,\ell}(\mathrm{Id})^2 = \mathrm{Var}\left(\sum_{q=1}^{p} t_U\left(F^q, W\right) \right). \tag{4.46}$$

Let $F, F' \in \mathcal{F}$ be two simple finite graphs, let $i \in V(F)$ and $i' \in V(F')$. We define a new graph $(F \bowtie F')(i, i') = (F \sqcup F')/\{i \sim i'\}$ which is the disjoint union of F and F' followed by a quotient where we identify the vertex i in $V(F)$ with the vertex i' in $V(F')$, see Figure 4.3.

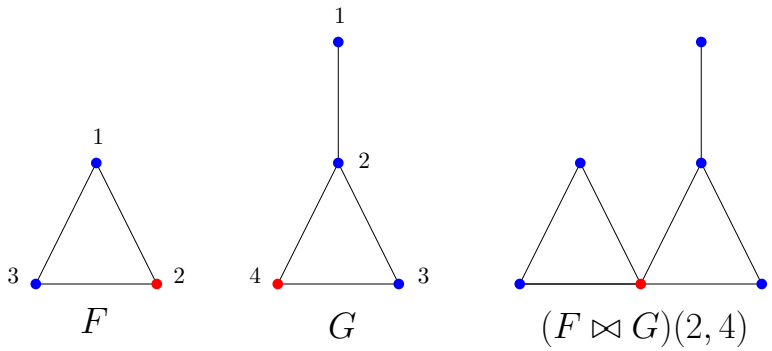


Figure 4.3 – Example of two graphs connected by two vertices.

With this notation, we have:

$$\begin{aligned}
\sigma^{F,\ell}(\mathrm{Id})^2 &= \mathbb{E}\left[\left(\sum_{q=1}^{p} t_U(F^q, W) \right)^2 \right] - \mathbb{E}\left[\sum_{q=1}^{p} t_U(F^q, W) \right]^2 \\
&= \sum_{q,q'=1}^{p} \int_0^1 t_x(F^q, W) t_x(F^{q'}, W) dx - \left(\sum_{q=1}^{p} \int_0^1 t_x(F^q, W) dx \right)^2 \\
&= \sum_{q,q'=1}^{p} t\big((F \bowtie F)(q, q'), W\big) - p^2\, t(F, W)^2.
\end{aligned} \tag{4.47}$$

Thus, we recover the limiting variance given in [84].

(iv) Let $d = 1$. We consider the two degenerate cases where no vertex is labelled ($k = 0$) or all vertices are labelled ($k = p$):

(a) for $k = 0$, we apply the δ-method to (4.46), to get that

$$\sqrt{n}\left[g(t_{\mathrm{inj}}(F, G_n)) - g(t(F, W))\right] \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^F(g)^2\right),$$

where

$$\sigma^F(g)^2 = g'(t(F, W))^2 \sigma^{F,\ell}(\mathrm{Id})^2. \tag{4.48}$$

(b) for $k = p$, we have $\Gamma_n^{F,\ell}(g) = (g(1) - g(0))t_{\mathrm{inj}}(F, G_n) + g(0)$, $\Gamma^{F,\ell}(g) = (g(1) - g(0))t(F^\ell, W)dx + g(0)$ and

$$\sigma^{F,\ell}(g)^2 = (g(1) - g(0))^2 \sigma^{F,\ell}(\mathrm{Id})^2. \tag{4.49}$$

(v) Let $d = 1$ and $F = K_2$. We have $\Gamma_n^{K_2,\ell}(\mathrm{Id}) = t(K_2, G_n)$ and we deduce from (4.46) that $\sigma^{K_2,\ell}(\mathrm{Id})^2 = 4\mathrm{Var}\ (D(U))$.

(a) If $k = 1$, then we have $\Gamma_n^{K_2^\bullet,\ell}(g) = \frac{1}{n}\sum_{i=1}^n g(D_i^{(n)})$ with $D_i^{(n)} = D_i(G_n)$ the normalized degree of i in G_n, see (4.23). We deduce from (4.44) that:

$$\sigma^{K_2^\bullet,\ell}(g)^2 = \mathrm{Var}\left(g(D(U)) + \int_0^1 W(x, U)g'(D(x))\,dx\right).$$

(b) If $k = 2$, using (iv)-(2), we get from (4.49) that:

$$\sigma^{K_2^{\bullet\bullet},\ell}(g)^2 = 4(g(1) - g(0))^2 \mathrm{Var}\ (D(U)),$$

where $K_2^{\bullet\bullet}$ denotes the complete graph K_2 with two labeled vertices.

(c) Finally, if $k = 0$, using (iv)-(1), we get from (4.48) that

$$\sigma^{K_2}(g)^2 = 4\,g'\left(\int_0^1 D(x)dx\right)^2 \mathrm{Var}\ (D(U)).$$

(vi) Thanks to (4.15), we get that Theorems 4.9 and 4.11 also hold with t_{inj} replaced by t.

(vii) It is left to reader to check that Theorem 4.11 is degenerate, that is $\sigma^{F,\ell}(g) = 0$, in the Erdős-Rényi case, that is $W \equiv \mathfrak{p}$ for some $0 \le \mathfrak{p} \le 1$, or when $\int_{[0,1]^k} \hat{t}_x(F, W)\,dx = 0$ and in particular when $t(F, W) = 0$.

The next corollary gives the limiting Gaussian process for the fluctuations of $(t_{\mathrm{inj}}(F, G_n) : F \in \mathcal{F})$.

Corollary 4.13. *We have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\,(t_{\mathrm{inj}}(F, G_n) - t(F, W)) : F \in \mathcal{F}\right) \xrightarrow[n\to+\infty]{(fdd)} \Theta_{\mathrm{inj}},$$

where $\Theta_{\mathrm{inj}} = (\Theta_{\mathrm{inj}}(F) : F \in \mathcal{F})$ is a centered Gaussian process with covariance function K_{inj} given, for $F, F' \in \mathcal{F}$, with $V(F) = [p]$ and $V(F') = [p']$, by:

$$K_{\mathrm{inj}}(F, F') = \mathrm{Cov}\left(\sum_{q=1}^p t_U(F^q, W), \sum_{q=1}^{p'} t_U\left(F'^q, W\right)\right) \tag{4.50}$$

$$= \sum_{q=1}^p \sum_{q'=1}^{p'} t\left((F \bowtie F')(q, q'), W\right) - pp'\ t(F, W)t(F', W). \tag{4.51}$$

Proof. We deduce from (4.45) and standard results on Gaussian vectors, the convergence, for the finite-dimensional distributions towards the Gaussian process with covariance function given by (4.50). Formula (4.51) can be derived similarly to (4.47). □

Remark 4.14. In particular, Corollary 4.13 proves a central limit theorem for quantum graphs (see Lovász [138], Section 6.1). A simple quantum graph is defined as a formal linear combination of a finite number of simple finite graphs with real coefficients. This definition makes it possible to study linear combination of homomorphism densities. For $F = (F_m : m \in [d]) \in \mathcal{F}^d$ and $a = (a_m : m \in [d]) \in \mathbb{R}^d$, we define the homomorphism density of $\mathfrak{F} = \sum_{m=1}^d a_m F_m$ for a graph G and a graphon $W \in \mathcal{W}$ as $t_{\mathrm{inj}}(\mathfrak{F}, G) = \langle a, t_{\mathrm{inj}}(F, G)\rangle$ and $t(\mathfrak{F}, W) = \langle a, t(F, W)\rangle$. We deduce from Corollary 4.13 the following convergence in distribution:

$$\sqrt{n}\left(t_{\mathrm{inj}}(\mathfrak{F}, G_n) - t(\mathfrak{F}, W)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma(\mathfrak{F})^2\right), \tag{4.52}$$

where $\sigma(\mathfrak{F})^2$ is given by (4.45) or equivalently $\sigma(\mathfrak{F})^2 = \sum_{m,m'\in[d]} a_m a_{m'} K_{\mathrm{inj}}(F_m, F_{m'})$.

We also have the limiting Gaussian process for the fluctuations of $(t_{\mathrm{ind}}(F, G_n) : F \in \mathcal{F})$.

Corollary 4.15. *We have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\left(t_{\mathrm{ind}}(F, G_n) - t_{\mathrm{ind}}(F, W)\right) : F \in \mathcal{F}\right) \xrightarrow[n\to+\infty]{(fdd)} \Theta_{\mathrm{ind}},$$

where $\Theta_{\mathrm{ind}} = (\Theta_{\mathrm{ind}}(F) : F \in \mathcal{F})$ is a centered Gaussian process with covariance function K_{ind} given, for $F_1, F_2 \in \mathcal{F}$, with $V(F_1) = [p_1]$ and $V(F_2) = [p_2]$, by:

$$\begin{aligned} &K_{\mathrm{ind}}(F_1, F_2) \\ &= \mathrm{Cov}\left(\sum_{F_1'\geq F_1} (-1)^{e(F_1')} \sum_{q=1}^{p_1} t_U\left((F_1')^q, W\right), \sum_{F_2'\geq F_2} (-1)^{e(F_2')} \sum_{q=1}^{p_2} t_U\left((F_2')^q, W\right)\right) \\ &= \sum_{F_1'\geq F_1,\, F_2'\geq F_2} (-1)^{e(F_1')+e(F_2')} \left(\sum_{q=1}^{p_1}\sum_{q'=1}^{p_2} t\left((F_1' \bowtie F_2')(q, q'), W\right) - p_1 p_2\ t(F_1', W) t(F_2', W)\right), \end{aligned} \tag{4.53}$$

where $F' \geq F$ means that $V(F) = V(F')$ and $E(F) \subset E(F')$, that is F' ranges over all simple graphs obtained from F by adding edges.

Proof. Notice that $t_{\mathrm{ind}}(F, G_n)$ is a linear combination of subgraph counts by Proposition 4.3. We deduce from (4.45) and standard results on Gaussian vectors, the convergence, for the finite-dimensional distributions towards the Gaussian process with covariance function given by the first equality in (4.53) which is derived from the second formula of (4.25) and (4.45). The second equality of (4.53) can be derived similarly to (4.51). □

4.4 A preliminary result

Let $F = (F_m : m \in [d]) \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ with $|\ell| = k$ such that Condition (4.38) holds. Let $W \in \mathcal{W}$ be a graphon and $X = (X_i : i \in \mathbb{N}^*)$ be a sequence of independent uniform random variables on $[0, 1]$. Let $n \in \mathbb{N}^*$ such that $n > p$. Let $G_n = G_n(W)$ be the associated W-random graphs with vertices $[n]$, see Section 4.2.4. Recall the definitions (4.13) of $Y^\beta(F, G)$, (4.18) of $\hat{Y}^\alpha(F^\ell, G^\alpha)$ and (4.19) of $\tilde{Y}^\beta(F^\ell, G^\alpha)$ for a simple finite graph F. We set $Y^\beta = (Y^\beta(F_m, G_n) : m \in [d])$ for $\beta \in \mathcal{S}_{n,p}$ and $\tilde{Y}^\beta = (\tilde{Y}^\beta(F_m^\ell, G_n^\alpha) : m \in [d])$ as well as $Y^\alpha = Y^\alpha(F_m^\ell, G_n^\alpha)$ (which does not

depend on $m \in [d]$) for $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$ and $\alpha \in \mathcal{S}_{p,k}$. Notice that for $\alpha \in \mathcal{S}_{p,k}$ and $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we have that, conditionally on X, $\hat{Y}^\alpha$ and $\tilde{Y}^\beta$ are independent, $\hat{Y}^\alpha$ is a Bernoulli random variable and:

$$Y^\beta = \hat{Y}^\alpha \, \tilde{Y}^\beta.$$

Recall that $t_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right) = \left(t_{\mathrm{inj}}\left(F_m^\ell, G_n^\alpha\right) : m \in [d]\right)$. With this notation, we get from equation (4.20) that for $\ell \in \mathcal{M}_p$ with $|\ell| = k$ and $\alpha \in \mathcal{S}_{n,k}$:

$$t_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right) = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} Y^\beta = \hat{Y}^\alpha \; \tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right), \tag{4.54}$$

with

$$\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right) = \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} \tilde{Y}^\beta. \tag{4.55}$$

We also set $Z^\beta = \mathbb{E}[Y^\beta | X]$ and $\tilde{Z}^\beta = \mathbb{E}[\tilde{Y}^\beta | X]$. Recall (4.28). We have, for $Z^\beta = (Z_m^\beta : m \in [d])$ and $\tilde{Z}^\beta = (\tilde{Z}_m^\beta : m \in [d])$ that for $m \in [d]$:

$$Z_m^\beta = \prod_{\{i,j\} \in E(F_m)} W(X_{\beta_i}, X_{\beta_j}) \quad \text{and} \quad \tilde{Z}_m^\beta = \prod_{\{i,j\} \in \tilde{E}(F_m^\ell)} W(X_{\beta_i}, X_{\beta_j}) = \tilde{Z}_m(X_\beta).$$

We recall that $\hat{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\hat{Y}^\alpha | \, X\right] = \mathbb{E}\left[\hat{Y}^\alpha | \, X_\alpha\right]$, see (4.34), to deduce that:

$$Z^\beta = \hat{t}_{X_\alpha}(F^\ell, W) \, \tilde{Z}^\beta. \tag{4.56}$$

Lemma 4.16. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ and $W \in \mathcal{W}$ be a graphon. Let $(M_\beta : \beta \in \mathcal{S}_{n,p})$ be a sequence of $\sigma(X)$-measurable $\mathbb{R}^d$-valued random variables and $n > p$. Assume Condition (4.38) holds and that there exists a finite constant K such that for all $\beta \in \mathcal{S}_{n,p}$, we have $\mathbb{E}\left[|M_\beta|^2\right] \leq K$. Then we have:*

$$\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \langle Y^\beta - Z^\beta, \, M_\beta \rangle\right)^2\right] \leq dK \frac{p(p-1)}{8n(n-1)}.$$

Proof. We first assume that $d = 1$. We denote by $\mathrm{Cov}(. | \, X)$ the conditional covariance given X. We have:

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \left(Y^\beta - Z^\beta\right) M_\beta\right)^2\right] \\
&\qquad = \frac{1}{|\mathcal{S}_{n,p}|^2} \sum_{\beta \in \mathcal{S}_{n,p}} \sum_{\gamma \in \mathcal{S}_{n,p}} \mathbb{E}\left[\mathbb{E}\left[\left(Y^\beta - \mathbb{E}[Y^\beta | X]\right)\left(Y^\gamma - \mathbb{E}[Y^\gamma | X]\right) M_\beta M_\gamma \middle| \, X\right]\right] \\
&\qquad = \frac{1}{|\mathcal{S}_{n,p}|^2} \sum_{\beta \in \mathcal{S}_{n,p}} \sum_{\gamma \in \mathcal{S}_{n,p}} \mathbb{E}\left[M_\beta M_\gamma \, \mathrm{Cov}(Y^\beta, Y^\gamma | \, X)\right] \\
&\qquad \leq \frac{1}{|\mathcal{S}_{n,p}|^2} \sum_{\beta \in \mathcal{S}_{n,p}} \sum_{\gamma \in \mathcal{S}_{n,p}} \mathbb{E}\left[|M_\beta M_\gamma| \, |\mathrm{Cov}(Y^\beta, Y^\gamma | \, X)|\right].
\end{aligned}$$

If the $[n]$-words β and γ have at most one character in common, that is $|\beta \bigcap \gamma| \leq 1$, then, by construction, Y^β and Y^γ are conditionally independent given X . This implies then that $\mathrm{Cov}(Y^\beta, Y^\gamma | \, X) = 0$. If $|\beta \bigcap \gamma| > 1$, then as Y^β and Y^γ are Bernoulli random variables

and we have the upper bound $|\mathrm{Cov}(Y^\beta, Y^\gamma | X)| \leq 1/4$. The number of possible choices for $\beta, \gamma \in \mathcal{S}_{n,p}$ such that $|\beta \bigcap \gamma| > 1$ is bounded from above by $A_n^p \binom{p}{2} A_{n-2}^{p-2}$. We deduce that:

$$\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,p}|}\sum_{\beta\in\mathcal{S}_{n,p}}\left(Y^\beta - Z^\beta\right)M_\beta\right)^2\right] \leq \frac{1}{4(A_n^p)^2} A_n^p \binom{p}{2} A_{n-2}^{p-2}\,\mathbb{E}\left[|M_\beta M_\gamma|\right]$$
$$\leq K\frac{p(p-1)}{8n(n-1)},$$

where we used the Cauchy-Schwarz inequality for the last inequality to get $\mathbb{E}\left[|M_\beta M_\gamma|\right] \leq K$.

In the case $d \geq 1$, the term $\mathbb{E}\left[|M_\beta M_\gamma|\right]$ in the above inequalities has to be replaced by $\mathbb{E}\left[|M_\beta|_1 |M_\gamma|_1\right]$, where $|\cdot|_1$ is the L^1 norm in $\mathbb{R}^d$. Then, use that $|x|_1^2 \leq d|x|^2$ and thus $\mathbb{E}\left[|M_\beta|_1 |M_\gamma|_1\right] \leq dK$ to conclude. □

The proof of the next Lemma is similar and left to the reader (notice the next lemma is in fact Lemma 4.16 stated for $d = 1$ and the graph $\mathcal{R}_\ell(F)_m$ of the labeled vertices, see Section 4.2.2 for the definition of $\mathcal{R}_\ell(F)_m$, which thanks to condition (4.38), does not depend on $m \in [d]$). We recall that $\hat{t}_{X_\alpha}(F^\ell, W) = \mathbb{E}\left[\hat{Y}^\alpha | X\right] = \mathbb{E}\left[\hat{Y}^\alpha | X_\alpha\right]$, see (4.34).

Lemma 4.17. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$ and $W \in \mathcal{W}$ be a graphon. Let $k \in [p]$ and $(M_\alpha : \alpha \in \mathcal{S}_{n,k})$ be a sequence of $\sigma(X)$-measurable $\mathbb{R}^d$-valued random variables and $n > p$. Assume Condition (4.38) holds and that there exists a finite constant K such that for all $\alpha \in \mathcal{S}_{n,k}$, we have $\mathbb{E}\left[|M_\alpha|^2\right] \leq K$. Then we have:*

$$\mathbb{E}\left[\left(\frac{1}{|\mathcal{S}_{n,k}|}\sum_{\alpha\in\mathcal{S}_{n,k}}\langle \hat{Y}^\alpha - \hat{t}_{X_\alpha}(F^\ell, W),\, M_\alpha\rangle\right)^2\right] \leq dK\frac{k(k-1)}{8n(n-1)}.$$

We also state a variant of Lemma 4.16, when working conditionally on X_α for some $\alpha \in \mathcal{S}_{n,k}$.

The next result is a key ingredient in the proof of Theorems 4.9 and 4.11. Recall $\tilde{t}_x\left(F^\ell, W\right) = \left(\tilde{t}_x\left(F_m^\ell, W\right) : m \in [d]\right)$ with $\tilde{t}_x$ defined in (4.29). Notice that for all $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$:

$$\tilde{t}_{X_\alpha}\left(F^\ell, W\right) = \mathbb{E}\left[\tilde{Z}^\beta | X_\alpha\right] = \mathbb{E}\left[\tilde{t}_{\mathrm{inj}}(F^\ell, G_n) | X_\alpha\right].$$

Lemma 4.18. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ with $k = |\ell|$, $\alpha \in \mathcal{S}_{n,k}$ and $W \in \mathcal{W}$ be a graphon. Assume Condition (4.38) holds. Then, we have:*

$$\mathbb{E}\left[\left|\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right) - \tilde{t}_{X_\alpha}\left(F^\ell, W\right)\right|^2 \,\middle|\, X_\alpha, \hat{Y}^\alpha\right] \leq d\frac{(p-k)}{4(n-k)}.$$

Proof. We consider the case $d = 1$. Recall the definition of $\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right)$ given in (4.55). Set:

$$\mathcal{A} = \mathbb{E}\left[\frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|^2}\left(\sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}}\left(\tilde{Y}^\beta - \tilde{t}_{X_\alpha}\left(F^\ell, W\right)\right)\right)^2 \,\middle|\, X_\alpha, \hat{Y}^\alpha\right].$$

Following the proof of Lemma 4.16 with $M_\beta = 1$, and using also that $\mathbb{E}\left[\tilde{Y}^\beta \,\middle|\, X_\alpha, \hat{Y}^\alpha\right] = \tilde{t}_{X_\alpha}\left(F^\ell, W\right)$, and that $\tilde{Y}^\beta$ and $\tilde{Y}^\gamma$ are conditionally on X_α independent of $\hat{Y}^\alpha$ for $\beta, \gamma \in \mathcal{S}_{n,p}^{\ell,\alpha}$, we get:

$$\mathcal{A} \leq \frac{1}{|\mathcal{S}_{n,p}^{\ell,\alpha}|^2}\sum_{\beta\in\mathcal{S}_{n,p}^{\ell,\alpha}}\sum_{\gamma\in\mathcal{S}_{n,p}^{\ell,\alpha}}|\mathrm{Cov}(\tilde{Y}^\beta, \tilde{Y}^\gamma | X_\alpha)|.$$

If β and γ have no more than α in common, that is $\beta \bigcap \gamma = \alpha$, then $\tilde{Y}^\beta$ and $\tilde{Y}^\gamma$ are conditionally on X_α independent and thus $\mathrm{Cov}(\tilde{Y}^\beta, \tilde{Y}^\gamma \,|\, X) = 0$.

If $|\beta \bigcap \gamma| > |\alpha|$, then as $\tilde{Y}^\beta$ and $\tilde{Y}^\gamma$ are Bernoulli random variables, we have the upper bound $|\mathrm{Cov}(\tilde{Y}^\beta, \tilde{Y}^\gamma \,|\, X)| \leq 1/4$. The number of possible choices for $\beta, \gamma \in \mathcal{S}_{n,p}^{\ell,\alpha}$ such that $|\beta \bigcap \gamma| > |\alpha|$ is bounded from above by $A_{n-k}^{p-k}(p-k)A_{n-k-1}^{p-k-1}$. We deduce that:

$$\mathcal{A} \leq \frac{1}{4(A_{n-k}^{p-k})^2} A_{n-k}^{p-k}(p-k)A_{n-k-1}^{p-k-1} \leq \frac{(p-k)}{4(n-k)}.$$

The extension to $d \geq 1$ is direct. □

4.5 Proof of Theorem 4.9

We first state a preliminary lemma.

Lemma 4.19. *Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$ with $k = |\ell|$, and $W \in \mathcal{W}$ be a graphon. Assume Condition (4.38) holds. Then, for all $n > k$ and $g \in \mathcal{C}^1([0,1])$, we have:*

$$\mathbb{E}\left[\left|\Gamma_n^{F,\ell}(g) - \hat{\Gamma}_n^{F,\ell}(g)\right|\right] \leq d \,\|g\|_\infty \sqrt{\frac{k(k-1)}{2n(n-1)}} + \frac{1}{2}\,\|\nabla g\|_\infty \sqrt{\frac{p-k}{n-k}}.$$

Proof. We first consider the case $d = 1$. Let $g \in \mathcal{C}^1([0,1])$. We first assume that $g(0) = 0$. Then, we deduce from the definition (4.40) of $\Gamma_n^{F,\ell}$ and from (4.54) and (4.55), as $\hat{Y}^\alpha \in \{0,1\}$, that:

$$\Gamma_n^{F,\ell}(g) = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha \, g\left(\tilde{t}_{\mathrm{inj}}\left(F^\ell, G_n^\alpha\right)\right).$$

And thus, using definition (4.41) of $\hat{\Gamma}_n^{F,\ell}$, we get $\left|\Gamma_n^{F,\ell}(g) - \hat{\Gamma}_n^{F,\ell}(g)\right| \leq B_1 + B_2$ with

$$B_1 = \frac{1}{|\mathcal{S}_{n,k}|} \Big| \sum_{\alpha \in \mathcal{S}_{n,k}} \left(\hat{Y}^\alpha - \hat{t}_{X_\alpha}(F^\ell, W)\right) g\left(\tilde{t}_{X_\alpha}(F^\ell, W)\right) \Big|$$

and

$$B_2 = \frac{1}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha \Big| g\left(\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) - g\left(\tilde{t}_{X_\alpha}(F^\ell, W)\right) \Big|.$$

Thanks to Lemma 4.17, we get $\mathbb{E}[B_1^2] \leq \|g\|_\infty^2 \, k(k-1)/8n(n-1)$. Thanks to Lemma 4.18, we get using Jensen's inequality that $\mathbb{E}[B_2^2] \leq \|g'\|_\infty^2 (p-k)/4(n-k)$. This gives the result when $g(0) = 0$, except there is a $1/2$ in front of $\|g\|_\infty$ in the upper bound of the Lemma. In general, use that $\Gamma_n^{F,\ell}$ and $\hat{\Gamma}_n^{F,\ell}$ are probability measures, so that $\left(\Gamma_n^{F,\ell} - \hat{\Gamma}_n^{F,\ell}\right)(g) = \left(\Gamma_n^{F,\ell} - \hat{\Gamma}_n^{F,\ell}\right)(\bar{g})$, with $\bar{g} = g - g(0)$. Then use that and $\|\bar{g}\|_\infty \leq 2\,\|g\|_\infty$ to conclude. The case $d \geq 1$ is similar. □

We can now prove Theorem 4.9.

Proof of Theorem 4.9. We first consider the case $d = 1$. Let $g \in \mathcal{C}^1([0,1])$. Using Lemma 4.19 and the first Borel-Cantelli lemma, we get that a.s. $\lim_{n\to\infty} \left(\Gamma_{\phi(n)}^{F,\ell}(g) - \hat{\Gamma}_{\phi(n)}^{F,\ell}(g)\right) = 0$, with $\phi(n) = n^4$. We notice that $\hat{\Gamma}_n^{F,\ell}(g)$ is a U-statistic with kernel $\Phi_1(X_{[k]})$ where for $x \in [0,1]^k$:

$$\Phi_1(x) = \hat{t}_x \, g\left(\tilde{t}_x\right) + \left(1 - \hat{t}_x\right) g(0),$$

with $t_x = t_x(F^\ell, W)$ and the obvious variants for $\tilde{t}_x$ and $\hat{t}_x$.

Moreover, because g is uniformly bounded on $[0,1]$, we get that Var $\left(\Phi_1(X_{[k]})\right) < +\infty$ and we can apply the law of large numbers for U-statistics to obtain that a.s. $\lim_{n\to\infty} \hat{\Gamma}_n^{F,\ell}(g) = \mathbb{E}[\Phi(X_{[k]})] = \Gamma^{F,\ell}(g)$. We deduce that a.s. $\lim_{n\to\infty} \Gamma_{\phi(n)}^{F,\ell}(g) = \Gamma^{F,\ell}(g)$.

Let $n' \geq n > k$. We have $\mathcal{S}_{n,k} \subset \mathcal{S}_{n',k}$ and $\mathcal{S}_{n,p}^{\ell,\alpha} \subset \mathcal{S}_{n',p}^{\ell,\alpha}$ for $\alpha \in \mathcal{S}_{n,k}$. Recall $|\mathcal{S}_{n,k}^{\ell,\alpha}| = A_{n-k}^{p-k}$. We deduce that for $\alpha \in \mathcal{S}_{n,k}$:

$$\begin{aligned}\left|t_{\mathrm{inj}}(F^\ell, G_n^\alpha) - t_{\mathrm{inj}}(F^\ell, G_{n'}^\alpha)\right| &\leq \frac{1}{A_{n'-k}^{p-k}}\left|A_{n'-k}^{p-k} - A_{n-k}^{p-k}\right| + \left|\frac{1}{A_{n'-k}^{p-k}} - \frac{1}{A_{n-k}^{p-k}}\right| A_{n-k}^{p-k} \\ &= 2\left(1 - \frac{A_{n-k}^{p-k}}{A_{n'-k}^{p-k}}\right) \\ &\leq 2\left(1 - \left(\frac{n-p}{n'-p}\right)^{p-k}\right).\end{aligned}$$

We deduce that:

$$\begin{aligned}\left|\Gamma_n^{F,\ell}(g) - \Gamma_{n'}^{F,\ell}(g)\right| \leq \frac{1}{A_{n'}^k}\left|A_{n'}^k - A_n^k\right| \|g\|_\infty + \left|\frac{1}{A_{n'}^k} - \frac{1}{A_n^k}\right| A_n^k \|g\|_\infty \\ + \frac{1}{|\mathcal{S}_{n,k}|}\sum_{\alpha\in\mathcal{S}_{n,k}} \left|g\left(t_{\mathrm{inj}}(F^\ell, G_n^\alpha)\right) - g\left(t_{\mathrm{inj}}(F^\ell, G_{n'}^\alpha)\right)\right| \\ \leq 2\|g\|_\infty\left(1 - \left(\frac{n-k}{n'-k}\right)^k\right) + 2\|g'\|_\infty\left(1 - \left(\frac{n-p}{n'-p}\right)^{p-k}\right).\end{aligned}$$

This implies that a.s. $\lim_{n\to\infty} \sup_{n'\in\{\phi(n),\ldots,\phi(n+1)\}} \left|\Gamma_{\phi(n)}^{F,\ell}(g) - \Gamma_{n'}^{F,\ell}(g)\right| = 0$.

With the first part of the proof, we deduce that for all $g \in \mathcal{C}^1([0,1])$, a.s. $\lim_{n\to\infty} \Gamma_n^{F,\ell}(g) = \Gamma^{F,\ell}(g)$. Since there exists a convergence determining countable subset of $\mathcal{C}^1([0,1])$, we get that a.s. $\lim_{n\to\infty} \Gamma_n^{F,\ell} = \Gamma^{F,\ell}$ for the weak convergence of the measures on $[0,1]$.

The proof for $d \geq 1$ is straightforward. □

4.6 Proof of Theorem 4.11

Let $\ell \in \mathcal{M}_p$ with $k = |\ell|$. We assume Condition (4.38) holds.

Recall the random probability measures $\Gamma_n^{F,\ell}$, $\hat{\Gamma}_n^{F,\ell}$ and $\Gamma^{F,\ell}$ are defined in (4.40), (4.41) and (4.42). Let $g \in \mathcal{C}^2([0,1]^d)$. We define the U-statistic

$$U_n(g) = \frac{1}{|\mathcal{S}_{n,p}|}\sum_{\beta\in\mathcal{S}_{n,p}} \Phi_2(X_\beta), \tag{4.57}$$

with kernel $\Phi_2(X_{[p]})$ given by, for $x \in [0,1]^p$:

$$\Phi_2(x) = \hat{t}_{x_\ell}\, g(\tilde{t}_{x_\ell}) + (1 - \hat{t}_{x_\ell})\, g(0) + \hat{t}_{x_\ell}\, \langle \nabla g(\tilde{t}_{x_\ell}), \tilde{Z}(x) - \tilde{t}_{x_\ell}\rangle, \tag{4.58}$$

with $\hat{t}_y = \hat{t}_y(F^\ell, W)$, $\tilde{t}_y = \tilde{t}_y(F^\ell, W)$ for $y \in [0,1]^k$ and $\tilde{Z}(x)$ defined in (4.28). Notice that:

$$\mathbb{E}[U_n(g)] = \Gamma^{F,\ell}(g). \tag{4.59}$$

We define the random signed measure $\Lambda_n^{F,\ell} = \sqrt{n}\left[\Gamma_n^{F,\ell} - \Gamma^{F,\ell}\right]$.

Lemma 4.20. *Let $W \in \mathcal{W}$ be a graphon. Let $F \in \mathcal{F}^d$ be a sequence of $d \geq 1$ simple finite graphs with $p = v(F)$, $\ell \in \mathcal{M}_p$, with $k = |\ell|$. Assume Condition (4.38) holds. Let $g \in \mathcal{C}^2([0,1]^d)$. Then, we have that $\lim_{n\to\infty} \Lambda_n^{F,\ell}(g) - \sqrt{n}\,(U_n(g) - \mathbb{E}[U_n(g)]) = 0$ in $L^1(\mathbb{P})$.*

Proof. Recall (4.55). We write:

$$\Lambda_n^{F,\ell}(g) - \sqrt{n}\,(U_n(g) - \mathbb{E}[U_n(g)]) = R_1(n) + R_2(n) + R_3(n) \tag{4.60}$$

with

$$R_1(n) = \frac{\sqrt{n}}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha\, H_1(\alpha),$$

$$R_2(n) = \frac{\sqrt{n}}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \left(\hat{Y}^\alpha - \hat{t}_{X_\alpha}\right) H_2(\alpha),$$

$$R_3(n) = \frac{\sqrt{n}}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \langle Y^\beta - Z^\beta,\, \nabla g(\tilde{t}_{X_{\beta_\ell}})\rangle$$

$$= \frac{\sqrt{n}}{|\mathcal{S}_{n,k}|} \sum_{\alpha \in \mathcal{S}_{n,k}} \hat{Y}^\alpha\, \langle \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha), \nabla g(\tilde{t}_{X_{\beta_\ell}})\rangle - \frac{\sqrt{n}}{|\mathcal{S}_{n,p}|} \sum_{\beta \in \mathcal{S}_{n,p}} \hat{t}_{X_{\beta_\ell}} \langle \tilde{Z}(X_\beta), \nabla g(\tilde{t}_{X_{\beta_\ell}})\rangle,$$

(where we used (4.54) and (4.56) for the last equality) and

$$H_1(\alpha) = g\big(\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha)\big) - g\big(\tilde{t}_{X_\alpha}\big) - \langle \tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) - \tilde{t}_{X_\alpha},\, \nabla g\big(\tilde{t}_{X_\alpha}\big)\rangle,$$
$$H_2(\alpha) = g\big(\tilde{t}_{X_\alpha}\big) - g(0) - \langle \tilde{t}_{X_\alpha}, \nabla g\big(\tilde{t}_{X_\alpha}\big)\rangle.$$

According to Lemma 4.16, we get that $\lim_{n\to\infty} R_3(n) = 0$ in $L^2(\mathbb{P})$. According to Lemma 4.17, and since $|H_2(\alpha)| \leq 2\,\|g\|_\infty + \|\nabla g\|_\infty$, we get that $\lim_{n\to\infty} R_2(n) = 0$ in $L^2(\mathbb{P})$. Since $g \in \mathcal{C}^2([0,1]^d)$, by Taylor-Lagrange inequality, we have that for all $x, y \in \mathbb{R}$,

$$|g(x) - g(y) - \langle x - y, \nabla g(x)\rangle| \leq \frac{1}{2}\,\|\nabla^2 g\|_\infty\, |x - y|^2.$$

This gives $|H_1(\alpha)| \leq \frac{1}{2}\,\|\nabla^2 g\|_\infty\, |\tilde{t}_{\mathrm{inj}}(F^\ell, G_n^\alpha) - \tilde{t}_{X_\alpha}|^2$. According to Lemma 4.18, we get that $\lim_{n\to\infty} R_1(n) = 0$ in $L^1(\mathbb{P})$. This ends the proof. □

We give a central limit theorem for the U-statistic U_n defined in (4.57).

Lemma 4.21. *Under the same hypothesis as in Lemma 4.20, we have the following convergence in distribution:*

$$\sqrt{n}\left(U_n(g) - \Gamma^{F,\ell}(g)\right) \xrightarrow[n\to+\infty]{(d)} \mathcal{N}\left(0, \sigma^{F,\ell}(g)^2\right),$$

with $\sigma^{F,\ell}(g)^2 = \mathrm{Var}\,(\mathcal{U})$ and, U being a uniform random variable on $[0,1]$:

$$\mathcal{U} = \sum_{i=1}^{k} \int_{[0,1]^k} \hat{t}_{R_i(x,U)}(F^\ell, W)\left(g\big(\tilde{t}_{R_i(x,U)}(F^\ell, W)\big) - g(0)\right) dx$$
$$+ \sum_{q \in [p]\setminus \ell} \int_{[0,1]^k} \langle \nabla g\big(\tilde{t}_x(F^\ell, W)\big), t_{xU}(F^{\ell q}, W)\rangle\, dx.$$

Proof. Recall the definition of $\tau_{ij}(\beta)$ given in (4.5). The random variable $U_n(g)$ is a U-statistic with bounded kernel. Since $\mathbb{E}[U_n(g)] = \Gamma^{F,\ell}(g)$, we deduce from the central limit theorem for U-statistics, see Theorem 7.1 in [105], that $\sqrt{n}\left(U_n(g) - \Gamma^{F,\ell}(g)\right)$ converges in distribution towards a centered Gaussian random variable with variance $\mathrm{Var}\,(\mathcal{U}')$ and $\mathcal{U}' = \sum_{q=1}^{p} \mathbb{E}\left[\Phi_2(\tau_{1q}(X))\,|\,X_1\right]$, and Φ_2 given by (4.58). We first compute $\mathbb{E}\left[\Phi_2(\tau_{1q}(X))\,|\,X_1\right]$ for $q \in [p]$. We distinguish the cases $q \notin \ell$ and $q \in \ell$.

The case $q \notin \{\ell_1, \dots, \ell_k\}$

Noticing that $\tau_{1q}(X)_\ell$ does not depend on X_1, we deduce that:

$$\begin{aligned}\mathbb{E}\left[\Phi_2(\tau_{1q}(X))\middle| X_1\right] &= \mathbb{E}\left[\hat{t}_{\tau_{1q}(X)_\ell}\, g\big(\tilde{t}_{\tau_{1q}(X)_\ell}\big) + \big(1-\hat{t}_{\tau_{1q}(X)_\ell}\big)\, g(0)\middle| X_1\right] \\ &\qquad + \mathbb{E}\left[\hat{t}_{\tau_{1q}(X)_\ell}\, \langle \nabla g\big(\tilde{t}_{\tau_{1q}(X)_\ell}\big), \tilde{Z}(\tau_{1q}(X)_{[p]}) - \tilde{t}_{\tau_{1q}(X)_\ell}\rangle\middle| X_1\right] \\ &= C + \int_{[0,1]^k} \hat{t}_x \langle \nabla g\big(\tilde{t}_x\big), \tilde{t}_{xX_1}(F^{\ell q}, W)\rangle\, dx \\ &= C + \int_{[0,1]^k} \langle \nabla g\big(\tilde{t}_x\big), t_{xX_1}(F^{\ell q}, W)\rangle\, dx,\end{aligned}$$

where C is a constant not depending on X_1 (which therefore will disappear when computing the variance of $\mathcal{U}'$).

The case $q \in \{\ell_1, \dots, \ell_k\}$

Let $q = \ell_i$ for some $i \in [k]$. Since $\mathbb{E}\left[\tilde{Z}(\tau_{1q}(X)_{[p]})\middle| \tau_{1q}(X)_\ell\right] = \tilde{t}_{\tau_{1q}(X)_\ell}$, we deduce that:

$$\begin{aligned}\mathbb{E}\left[\Phi_2(\tau_{1q}(X))\middle| X_1\right] &= \mathbb{E}\left[\hat{t}_{\tau_{1q}(X)_\ell}\, g\big(\tilde{t}_{\tau_{1q}(X)_\ell}\big) + \big(1-\hat{t}_{\tau_{1q}(X)_\ell}\big)\, g(0)\middle| X_1\right] \\ &= g(0) + \int_{[0,1]^k} \hat{t}_{R_i(x,X_1)} \Big(g\big(\tilde{t}_{R_i(x,X_1)}\big) - g(0)\Big)\, dx.\end{aligned}$$

Thus, we obtain that $\mathcal{U}' = \mathcal{U} + C'$ for some constant C' and:

$$\mathcal{U} = \sum_{i=1}^{k} \int_{[0,1]^k} \hat{t}_{R_i(x,X_1)} \Big(g\big(\tilde{t}_{R_i(x,X_1)}\big) - g(0)\Big)\, dx + \sum_{q \notin \ell} \int_{[0,1]^k} \langle \nabla g\big(\tilde{t}_x\big), t_{xX_1}(F^{\ell q}, W)\rangle\, dx.$$

This gives the result. □

Theorem 4.11 is then a direct consequence of Lemmas 4.20 and 4.21 and (4.59).

4.7 Asymptotics for the empirical degrees cumulative distribution function of the degrees

Let W be a graphon on $[0,1]$ and $n \in \mathbb{N}^*$. Recall the definition of the normalized degree function D of the graphon W given in (4.33), $D(x) = \int_{[0,1]} W(x,y)dy = t_x(K_2^\bullet, W)$. From Section 4.2.4, recall that $G_n = G_n(W)$ is the associated W-random graph with n vertices constructed from W and the sequence $X = (X_i : i \in \mathbb{N}^*)$ of independent uniform random variables on $[0,1]$. Recall the (normalized) degree sequence of a graph defined in (4.23), and set

$$D_i^{(n)} = D_i(G_n) = t_{\mathrm{inj}}(K_2^\bullet, G_n^i)$$

the normalized degree of the vertex $i \in [n]$ in G_n. By construction of G_n, we get that conditionally on X_i , $(n-1)D_i^{(n)}$ is for $n \geq i$ a binomial random variable with parameters $(n-1, D(X_i))$. We define the empirical cumulative distribution function $\Pi_n = (\Pi_n(y) : y \in [0,1])$ of the degrees of the graph G_n by, for $y \in [0,1]$:

$$\Pi_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\left\{D_i^{(n)} \leq D(y)\right\}}. \tag{4.61}$$

Remark 4.22. If we take $g = \mathbf{1}_{[0,D(y)]}$ with $y \in [0,1]$ and $F = K_2$ in (4.3) and using the expression of $\Gamma^{F,\ell}$ given in Remark 4.8, (ii), we have that $\Pi_n(y) = \Gamma_n^{K_2,\bullet}(g)$ and $\Gamma^{K_2,\bullet}(g) = y$. If D is increasing, then $\Gamma^{K_2,\bullet}$, which is the distribution of $D(U)$, with U uniform on $[0,1]$, has no atoms and Theorem 4.9 implies that a.s. $\lim_{n\to\infty} \Pi_n(y) = y$ for all $y \in [0,1]$. Using Dini's theorem, we get that if D is increasing on $[0,1]$, then the function Π_n converges almost surely towards Id, the identity map on $[0,1]$, with respect to the uniform norm.

To get the corresponding fluctuations, we shall consider the following conditions:

$$W \in \mathcal{C}^3([0,1]^2),\ D' > 0,\ W \le 1 - \varepsilon_0 \text{ and } D \ge \varepsilon_0 \text{ for some } \varepsilon_0 \in (0, 1/2)\,. \tag{4.62}$$

If (4.62) holds, then we have $D \in \mathcal{C}^1([0,1])$ and $D([0,1]) \subset [\varepsilon_0, 1-\varepsilon_0]$. Notice that even if (4.62) holds, the set $\{W = 0\}$ might have positive Lebesgue measure; but the regularity conditions on W rules out bipartite graphons (but not tripartite graphons).

Theorem 4.23. *Assume that W satisfies condition (4.62). Then we have the following convergence of finite-dimensional distributions:*

$$\left(\sqrt{n}\,(\Pi_n(y) - y) : y \in (0,1)\right) \xrightarrow[n\to+\infty]{(fdd)} \chi,$$

where $\chi = (\chi_y : y \in (0,1))$ is a centered Gaussian process defined, for all $y \in (0,1)$ by:

$$\chi_y = \int_0^1 (\rho(y,u) - \bar\rho(y))dB_u, \tag{4.63}$$

with $B = (B_u, u \ge 0)$ a standard Brownian motion, and $(\rho(y,u) : u \in [0,1])$ and $\bar\rho(y)$ defined for $y \in (0,1)$ by:

$$\rho(y,u) = \mathbf{1}_{[0,y]}(u) - \frac{W(y,u)}{D'(y)} \quad \text{and} \quad \bar\rho(y) = \int_0^1 \rho(y,u)du.$$

Remark 4.24. The covariance kernel of the Gaussian process χ can also be written as $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$, where for $y, z \in (0,1)$:

$$\Sigma_1(y,z) = y \wedge z - yz, \tag{4.64}$$

$$\Sigma_2(y,z) = \frac{1}{D'(y)D'(z)}\left(\int_0^1 W(y,x)W(z,x)dx - D(y)D(z)\right), \tag{4.65}$$

$$\Sigma_3(y,z) = \frac{1}{D'(y)}\left(D(y)z - \int_0^z W(y,x)dx\right) + \frac{1}{D'(z)}\left(D(z)y - \int_0^y W(z,x)dx\right). \tag{4.66}$$

Thus, for $y \in (0,1)$ the variance of $\chi(y)$ is:

$$\Sigma(y,y) = y(1-y) + \frac{1}{D'(y)^2}\left(\int_0^1 W(y,x)^2dx - D(y)^2\right) + \frac{2}{D'(y)}\left(D(y)y - \int_0^y W(y,x)dx\right).$$

Remark 4.25. We conjecture that the convergence of Theorem 4.23 holds for the process in the Skorokhod space. However, the techniques used to prove this theorem are not strong enough to get such a result.

4.8 Preliminary results for the empirical cdf of the degrees

4.8.1 Estimates for the first moment of the empirical cdf

Recall $X = (X_n : n \in \mathbb{N}^*)$ is a sequence of independent random variables uniformly distributed on $[0,1]$ used to construct the sequence of W-random graphs $(G_n : n \in \mathbb{N}^*)$. Recall $\Pi_n(y)$ is given in (4.61).

For all $y \in (0,1)$, we set $c_n(y) = \mathbb{E}\left[\Pi_{n+1}(y)\right]$ that is

$$c_n(y) = \mathbb{P}\left(D_1^{(n+1)} \leq D(y)\right), \tag{4.67}$$

where, conditionally on X, $D_1^{(n+1)}$ is a binomial random variable with parameter $(n, D(X_1))$. We set:

$$\sigma_{(x)}^2 = x(1-x) \quad \text{for } x \in [0,1], \tag{4.68}$$

and with $\lceil x \rceil$ the unique integer such that $\lceil x \rceil - 1 < x \leq \lceil x \rceil$,

$$S(x) = \lceil x \rceil - x - \frac{1}{2} \quad \text{for } x \in \mathbb{R}. \tag{4.69}$$

The next proposition gives precise asymptotics of c_n.

Proposition 4.26. *Assume that W satisfies condition (4.62). For all $y \in (0,1)$, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}^*$, we have with $d = D(y)$,*

$$n\left(c_n(y) - y\right) = -\frac{D''(y)}{D'(y)^3}\frac{\sigma_{(d)}^2}{2} + \frac{1}{D'(y)}\left(\frac{1-2d}{2} + S(nd)\right) + R_n^{4.26},$$

with

$$\left|R_n^{4.26}\right| \leq C\, n^{-\frac{1}{4}}.$$

In particular, because $|S(x)| \leq \frac{1}{2}$, for all $x \in \mathbb{R}$, we have that for all $y \in (0,1)$:

$$c_n(y) - y = O\left(n^{-1}\right). \tag{4.70}$$

Proof. For $n \in \mathbb{N}^*$, $d \in [0,1]$, $\delta \in \mathbb{R}$ and $\mathfrak{p} \in (0,1)$, we consider the CDF:

$$\mathcal{H}_{n,d,\delta}(\mathfrak{p}) = \mathbb{P}\left(X \leq nd + \delta\right),$$

where X a binomial random variable with parameters $(n, \mathfrak{p})$. Let $y \in (0,1)$. We have:

$$c_n(y) - y = \int_0^1 \left(\mathcal{H}_{n,d,0}(D(x)) - \mathbf{1}_{\{x \leq y\}}\right) dx.$$

By Proposition 4.40 applied with $G(x) = 1$ and $\delta = 0$, we obtain that:

$$n\int_0^1 \left(\mathcal{H}_{n,d,0}\left(D(x)\right) - \mathbf{1}_{\{x \leq y\}}\right) dx = -\frac{D''(y)}{D'(y)^3}\frac{\sigma_{(d)}^2}{2} + \frac{1}{D'(y)}\left(\frac{1-2d}{2} + S(nd)\right) + R_n^{4.26},$$

with $R_n^{4.26} = R_n^{4.40}(1)$ and $\left|R_n^{4.26}\right| \leq Cn^{-\frac{1}{4}}$. □

For $y \in (0,1)$ and $u \in [0,1]$, we set, with $d = D(y)$,

$$H_n(y,u) = n\left(\mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}} \middle| X_2 = u\right] - c_n(y)\right) \tag{4.71}$$

and

$$H_n^\star(y,u) = \mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}} \middle| X_1 = u\right] - c_n(y). \tag{4.72}$$

Proposition 4.27. *Assume that W satisfies condition (4.62). For all $y \in (0,1)$, there exists a positive constant C such that for all $n \geq 2$ and $u \in [0,1]$, we have with $d = D(y)$:*

$$H_n(y,u) = \frac{1}{D'(y)}\left(d - W(y,u)\right) + R_n^{4.27}(u),$$

with

$$\left|R_n^{4.27}(u)\right| \leq C\, n^{-\frac{1}{4}}. \tag{4.73}$$

For all $y \in (0,1)$ and $u \in [0,1]$ such that $u \neq y$, we have

$$|H_n^\star(y,u)| \leq 1 \quad \text{for all } n \geq 2, \text{ and} \quad \lim_{n\to\infty} H_n^\star(y,u) = \mathbf{1}_{\{u \leq y\}} - y. \tag{4.74}$$

Proof. In what follows, C denotes a positive constant which depends on ε_0, W and $y \in (0,1)$, and it may vary from line to line. Recall that $X_{[2]} = (X_1, X_2)$. We define the function φ_n by:

$$\varphi_n(x,u) = \mathbb{P}\left(D_1^{(n+1)} \leq d \,\middle|\, X_{[2]} = (x,u)\right) - \mathbf{1}_{\{x \leq y\}} \quad \text{for } x,u \in [0,1]. \tag{4.75}$$

Then we have for $u \in [0,1]$:

$$H_n(y,u) = n\mathbb{E}\left[\varphi_n(X_1,u)\right] - n(c_n(y) - y). \tag{4.76}$$

Conditionally on $\{X_{[2]} = (x,u)\}$, $D_1^{(n+1)}$ is distributed as $Y_{12} + \tilde{B}^{(n)}$, where Y_{12} and $\tilde{B}$ are independent, Y_{12} is Bernoulli $W(x,u)$ and $\tilde{B}$ is binomial with parameter $(n-1, D(x))$. Thus, we have:

$$\begin{aligned}
\varphi_n(x,u) &= \mathbb{P}\left(Y_{12} + \tilde{B} \leq nd\right) - \mathbf{1}_{\{x \leq y\}} \\
&= W(x,u)\left[\mathbb{P}\left(\tilde{B} \leq nd - 1\right) - \mathbf{1}_{\{x \leq y\}}\right] + (1 - W(x,u))\left[\mathbb{P}\left(\tilde{B} \leq nd\right) - \mathbf{1}_{\{x \leq y\}}\right] \\
&= W(x,u)\left[\mathcal{H}_{n-1,d,d-1}(D(x)) - \mathbf{1}_{\{x \leq y\}}\right] + (1 - W(x,u))\left[\mathcal{H}_{n-1,d,d}(D(x)) - \mathbf{1}_{\{x \leq y\}}\right].
\end{aligned} \tag{4.77}$$

Let $W_1(x,u)$ denote $\partial W(x,u)/\partial x$. We apply Proposition 4.40 with $G(x) = W(x,u)$, $\delta = d-1$ and n replaced by $n-1$ to get that:

$$\begin{aligned}
(n-1)\mathbb{E}\left[W(X_1,u)\left[\mathcal{H}_{n-1,d,d-1}(D(X_1)) - \mathbf{1}_{\{X_1 \leq y\}}\right]\right] \\
= \frac{\sigma_{(d)}^2}{2D'(y)^2}\left[W_1(y,u) - \frac{W(y,u)D''(y)}{D'(y)}\right] \\
+ \frac{W(y,u)}{D'(y)}\left(-\frac{1}{2} + S(nd-1)\right) + R_{n-1}^{4.40}(W(.,u)),
\end{aligned} \tag{4.78}$$

and with $G(x) = 1 - W(x,u)$, $\delta = d$ and n replaced by $n-1$, to get that:

$$\begin{aligned}
(n-1)\mathbb{E}\left[(1 - W(X_1,u))\left[\mathcal{H}_{n-1,d,d}(D(X_1)) - \mathbf{1}_{\{X_1 \leq y\}}\right]\right] \\
= \frac{\sigma_{(d)}^2}{2D'(y)^2}\left[-W_1(y,u) - \frac{(1 - W(y,u))D''(y)}{D'(y)}\right] \\
+ \frac{1 - W(y,u)}{D'(y)}\left(\frac{1}{2} + S(nd)\right) + R_{n-1}^{4.40}(1 - W(.,u)).
\end{aligned} \tag{4.79}$$

By equations (4.77), (4.78) and (4.79) and since $S(nd-1) = S(nd)$, we get that:

$$(n-1)\mathbb{E}\left[\varphi_n(X_1,u)\right] = -\frac{\sigma_{(d)}^2}{2}\frac{D''(y)}{D'(y)^3} + \frac{1}{D'(y)}\left(\frac{1}{2} - W(y,u) + S(nd)\right) + R_n^{(1)}(u),$$

where $R_n^{(1)}(u) = R_{n-1}^{4.40}(W(.,u)) + R_{n-1}^{4.40}(1 - W(.,u))$. Because W satisfies condition (4.62), we deduce from (4.122) that $\left|R_n^{(1)}(u)\right| \leq Cn^{-1/4}$ for some finite constant C which does not depend on n and $u \in [0,1]$. Using Proposition 4.26, we get that

$$(n-1)\mathbb{E}\left[\varphi_n(X_1,u)\right] - (n-1)(c_n(y) - y) = \frac{d - W(y,u)}{D'(y)} + R_n^{(2)}(u), \tag{4.80}$$

where $R_n^{(2)}(u) = R_n^{(1)}(u) + R_n^{4.26} + (c_n(y) - y)$ and $\left|R_n^{(2)}(u)\right| \leq Cn^{-1/4}$ because of (4.70). By equations (4.76) and (4.80), we deduce that:

$$H_n(y,u) = \frac{n}{n-1}\frac{d - W(y,u)}{D'(y)} + \frac{n}{n-1}R_n^{(2)}(u) = \frac{d - W(y,u)}{D'(y)} + R_n^{4.27}(u),$$

with $\left|R_n^{4.27}(y,u)\right| \leq Cn^{-\frac{1}{4}}$. This gives (4.73).

For the second assertion (4.74), we notice that

$$H_n^\star(y,u) = \mathcal{H}_{n,d,0}(D(u)) - c_n(y),$$

with $\mathcal{H}_{n,d,0}(D(u)) \in [0,1]$ and $c_n(y) \in [0,1]$. By the strong law of large numbers, we have for $u \neq y$:

$$\lim_{n\to\infty} \mathcal{H}_{n,d,0}(D(u)) = \mathbf{1}_{\{u\leq y\}}.$$

Using (4.70), we get the expected result. □

4.8.2 Estimates for the second moment of the empirical cdf

For $\mathrm{y} = (y_1, y_2) \in [0,1]^2$, let $M(\mathrm{y})$ be the covariance matrix of a pair (Y_1, Y_2) of Bernoulli random variables such that $\mathbb{P}(Y_i = 1) = D(y_i)$ for $i \in \{1,2\}$ and $\mathbb{P}(Y_1 = Y_2 = 1) = \int_{[0,1]} W(y_1,z)W(y_2,z)\,dz$.

Let $\mathcal{K}$ be the set of all convex sets in $\mathbb{R}^2$. For $K \in \mathcal{K}$, we define its sum with a vector x in $\mathbb{R}^2$ as

$$K + \mathrm{x} = \{k + \mathrm{x} : k \in K\}$$

and its product with a real matrix A of size 2×2 as

$$AK = \{Ak : k \in K\}.$$

Recall that for $\mathrm{x} \in \mathbb{R}^2$, $|\mathrm{x}|$ is the Euclidean norm of x in $\mathbb{R}^2$. Recall $X_{[2]} = (X_1, X_2)$. We define $\mathcal{D}^{(n+1)} = (\mathcal{D}_1^{(n+1)}, \mathcal{D}_2^{(n+1)})$, where for $i \in \{1,2\}$, $\mathcal{D}_i^{(n+1)}$ is the number of edges from the vertex i to the vertices $\{k, 3 \leq k \leq n+1\}$ of G_{n+1}; it is equal to $nD_i^{(n+1)}$ if the edge $\{1,2\}$ does not belong to G_{n+1} and to $nD_i^{(n+1)} - 1$ otherwise. The proof of the next proposition is postponed to section 4.11.

Proposition 4.28. *Assume that W satisfies condition (4.62). There exists a finite constant C_0 such that for all $\mathrm{x} = (x_1, x_2) \in [0,1]^2$ with $x_1 \neq x_2$, we get for all $n \geq 2$:*

$$\sup_{K\in\mathcal{K}} \left| \mathbb{P}\left(\mathcal{D}^{n+1} \in K \,|\, X_{[2]} = \mathrm{x}\right) - \mathbb{P}\left(Z \in \frac{M(\mathrm{x})^{-\frac{1}{2}}}{\sqrt{n-1}}(K - \mu(\mathrm{x}))\right)\right| \leq \frac{C_0}{\sqrt{n}},$$

where $\mu(\mathrm{x}) = (n-1)(D(x_1), D(x_2))$ and Z is a standard 2-dimensional Gaussian vector.

For $y_1, y_2 \in (0,1)$, with $d_1 = D(y_1)$ and $d_2 = D(y_2)$, we set with $\delta \in \mathbb{R}$ and $\mathrm{x} = (x_1, x_2) \in [0,1]^2$ such that $x_1 \neq x_2$:

$$\Psi_{n,\delta}(\mathrm{x}) = \mathbb{E}\left[\prod_{i\in\{1,2\}} \left(\mathbf{1}_{\{\mathcal{D}_i^{(n+1)} \leq nd_i + \delta\}} - \mathbf{1}_{\{X_i \leq y_i\}}\right) \Big| X_{[2]} = \mathrm{x}\right].$$

Recall Σ_2 defined in (4.65) and that $X_{[2]} = (X_1, X_2)$.

Lemma 4.29. *Assume that W satisfies condition (4.62). For all $\mathrm{y} = (y_1, y_2) \in (0,1)^2$, $\delta \in [-1, 0]$ and $G \in \mathcal{C}^1([0,1]^2)$, we have:*

$$\lim_{n\to\infty} n\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}\left(X_{[2]}\right)\right] = G(\mathrm{y})\Sigma_2(\mathrm{y}).$$

Proof. Let $A = 4\sqrt{\log(n-1)}$. For $n \geq 2$ and $\delta \in [-1,0]$, we set:

$$\Psi_{n,\delta}^{(1)}(\mathrm{x}) = \Psi_{n,\delta}(\mathrm{x}) \prod_{i=1}^{2} \mathbf{1}_{\{\sqrt{n-1}|D(x_i) - d_i| \leq A\}} \quad \text{and} \quad \Psi_{n,\delta}^{(2)}(\mathrm{x}) = \Psi_{n,\delta}(\mathrm{x}) - \Psi_{n,\delta}^{(1)}(\mathrm{x}).$$

Then we have

$$\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}\left(X_{[2]}\right)\right] = \sum_{i\in\{1,2\}} \mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(i)}\left(X_{[2]}\right)\right]. \tag{4.81}$$

Study of $\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(2)}\left(X_{[2]}\right)\right]$

Recall that for $i\in\{1,2\}$, conditionally on $X_i=x_i$, $\mathcal{D}_i^{(n+1)}$ is distributed as a Bernoulli random variable with parameter $(n-1,d_i)$. We get that:

$$\begin{aligned}\left|\Psi_{n,\delta}^{(2)}(\mathrm{x})\right| &\le 2\sum_{i\in\{1,2\}}\mathbb{E}\left[\left|\mathbf{1}_{\{\mathcal{D}_i^{(n+1)}\le nd_i+\delta\}}-\mathbf{1}_{\{X_i\le y_i\}}\right|\mathbf{1}_{\{\sqrt{n-1}\,|D(x_i)-d_i|\ge A\}}\,\middle|\,X_{[2]}=\mathrm{x}\right]\\ &= 2\sum_{i\in\{1,2\}}\left|\mathcal{H}_{n-1,d_i,\delta+d_i}(x_i)-\mathbf{1}_{\{x_i\le y_i\}}\right|\mathbf{1}_{\{\sqrt{n-1}\,|D(x_i)-d_i|\ge A\}}.\end{aligned}$$

By Lemma 4.39 (with n replaced by $n-1$), we deduce that:

$$\lim_{n\to\infty} n\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(2)}\left(X_{[2]}\right)\right]=0. \tag{4.82}$$

Study of $\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(1)}\left(X_{[2]}\right)\right]$

This part is more delicate. For $\mathfrak{z}=(z_1,z_2)\in[0,1]^2$, set $H(\mathfrak{z})=\frac{G(\mathfrak{z})}{D'(z_1)D'(z_2)}$ and $\mathfrak{t}_n(\mathfrak{z})=(t_n(z_1),t_n(z_2))$:

$$t_n(z_i)=D^{-1}\left(d_i+\frac{z_i}{\sqrt{n-1}}\right)\quad\text{for } i\in\{1,2\}.$$

Using the change of variable $z_i=\sqrt{n-1}(D(x_i)-d_i)$ for $i\in\{1,2\}$ with $\mathrm{x}=(x_1,x_2)$, we get:

$$\begin{aligned}(n-1)\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(1)}\left(X_{[2]}\right)\right] &= (n-1)\int_{[0,1]^2} G(\mathrm{x})\Psi_{n,\delta}(\mathrm{x})\prod_{i\in\{1,2\}}\mathbf{1}_{\{\sqrt{n-1}|D(x_i)-d_i|\le A\}}d\mathrm{x}\\ &= \int_{[-A,A]^2} H\left(\mathfrak{t}_n(\mathfrak{z})\right)\Psi_{n,\delta}\left(\mathfrak{t}_n(\mathfrak{z})\right)d\mathfrak{z}.\end{aligned} \tag{4.83}$$

Notice that:

$$\Psi_{n,\delta}\left(\mathfrak{t}_n(\mathfrak{z})\right)=\mathbb{E}\left[\prod_{i\in\{1,2\}}\left(\mathbf{1}_{\{\mathcal{D}_i^{(n+1)}\le nd_i+\delta\}}-\mathbf{1}_{\{z_i\le 0\}}\right)\,\middle|\,X_{[2]}=\mathfrak{t}_n(\mathfrak{z})\right].$$

Set $\tilde{\delta}=(\delta,\delta)$. We define the sets for $\mathfrak{z}$ and $\mathcal{D}^{(n+1)}$:

$$\begin{aligned} I^{(1)}&=[0,A]^2 \quad\text{and}\quad \tilde{C}_n^{(1)}=\tilde{\delta}+n\,(-\infty,d_1]\times(-\infty,d_2],\\ I^{(2)}&=[0,A]\times[-A,0) \quad\text{and}\quad \tilde{C}_n^{(2)}=\tilde{\delta}+n\,(-\infty,d_1]\times(d_2,+\infty),\\ I^{(3)}&=[-A,0)\times[0,A] \quad\text{and}\quad \tilde{C}_n^{(3)}=\tilde{\delta}+n\,(d_1,+\infty)\times(-\infty,d_2],\\ I^{(4)}&=[-A,0)^2 \quad\text{and}\quad \tilde{C}_n^{(4)}=\tilde{\delta}+n\,(d_1,+\infty)\times(d_2,+\infty).\end{aligned}$$

For $1\le i\le 4$, we set:

$$Q_n^{(i)}(\mathfrak{z})=\mathbb{P}\left(\mathcal{D}^{(n+1)}\in\tilde{C}_n^{(i)}\middle|X_{[2]}=\mathfrak{t}_n(\mathfrak{z})\right)\quad\text{and}\quad\Delta_n^{(i)}=\int_{I^{(i)}}H\left(\mathfrak{t}_n(\mathfrak{z})\right)Q_n^{(i)}(\mathfrak{z})\,d\mathfrak{z}.$$

By construction, we have:

$$(n-1)\mathbb{E}\left[G\left(X_{[2]}\right)\Psi_{n,\delta}^{(1)}\left(X_{[2]}\right)\right]=\sum_{i=1}^{4}\Delta_n^{(i)}. \tag{4.84}$$

We now study $\Delta_n^{(1)}$. By Proposition 4.28, we get that

$$\Delta_n^{(1)} = \int_{[0,A]^2} H\left(\mathfrak{t}_n(\mathfrak{z})\right) \mathbb{P}\left(Z \in \frac{M\left(\mathfrak{t}_n(\mathfrak{z})\right)^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right) d\mathfrak{z} + R_n^{(1)}$$

where $\mu(\mathrm{x}) = (n-1)(D(x_1), D(x_2))$ and $|R_n^{(1)}| \leq \|H\|_\infty 8C_0\sqrt{\log(n)/n}$ so that $\lim_{n\to\infty} R_n^{(1)} = 0$. Set $\tilde{d} = (d_1, d_2)$. Since $\mathfrak{t}_n(\mathfrak{z})$ converges towards y, we get:

$$\lim_{n\to\infty} H(\mathfrak{t}_n(\mathfrak{z})) = H(\mathrm{y}) \quad \text{and} \quad \lim_{n\to\infty} M\left(\mathfrak{t}_n(\mathfrak{z})\right) = M(\mathrm{y})$$

and, with $J(\mathfrak{z}) = (-\infty, -z_1] \times (-\infty, -z_2]$,

$$\frac{1}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right) = (n-1)^{-1/2}(\tilde{\delta} + \tilde{d}) + J(\mathfrak{z}). \tag{4.85}$$

Since $\lim_{n\to\infty} M\left(\mathfrak{t}_n(\mathfrak{z})\right) = M(\mathrm{y})$ and $M(\mathrm{y})$ is positive definite, we deduce that $d\mathrm{x}$-a.e.:

$$\lim_{n\to\infty} \mathbf{1}_{\frac{M(\mathfrak{t}_n(\mathfrak{z}))^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)}(\mathrm{x}) = \mathbf{1}_{M(\mathrm{y})^{-1/2} J(\mathfrak{z})}(\mathrm{x})$$

and thus (by dominated convergence):

$$\lim_{n\to\infty} \mathbb{P}\left(Z \in \frac{M\left(\mathfrak{t}_n(\mathfrak{z})\right)^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right) = \mathbb{P}\left(M(\mathrm{y})^{1/2} Z \in J(\mathfrak{z})\right).$$

For $\mathfrak{z} \in [2, +\infty)^2$ and $n \geq 2$, we have:

$$\begin{aligned}\left\{Z \in \frac{M\left(\mathfrak{t}_n(\mathfrak{z})\right)^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right\} &\subset \left\{2M\left(\mathfrak{t}_n(\mathfrak{z})\right)^{1/2} Z \in J(\mathfrak{z})\right\} \\ &\subset \left\{2^{3/2}|Z| \geq |\mathfrak{z}|\right\},\end{aligned}$$

where we used (4.85) and that for all $i \in \{1, 2\}$, $|(n-1)^{-1/2}(\delta + d_i)| \leq z_i/2$ for the first inclusion and for the second that $|M\mathrm{x}| \leq \sqrt{2}\,\|M\|_\infty\,|\mathrm{x}|$, $\|M^{1/2}\|_\infty \leq \sqrt{2}\,\|M\|_\infty^{1/2}$ and $\|M(\mathrm{x})\|_\infty \leq 1/2$ for all $\mathrm{x} \in [0,1]^2$ so that $|M\left(\mathfrak{t}_n(\mathfrak{z})\right)^{1/2} Z| \leq \sqrt{2}\,|Z|$. Since $\int_{\mathbb{R}} \mathbb{P}(2^{3/2}|Z| \geq |\mathfrak{z}|)\, d\mathfrak{z}$ is finite and H is bounded, we deduce from dominated convergence that:

$$\begin{aligned}\lim_{n\to\infty} \int_{[0,A]^2} H\left(\mathfrak{t}_n(\mathfrak{z})\right) \mathbb{P}\left(Z \in \frac{M\left(\mathfrak{t}_n(\mathfrak{z})\right)^{-1/2}}{\sqrt{n-1}} \left(\tilde{C}_n^{(1)} - \mu(\mathfrak{t}_n(\mathfrak{z}))\right)\right) d\mathfrak{z} \\ = H(\mathrm{y}) \int_{[0,+\infty)^2} \mathbb{P}\left(M(\mathrm{y})^{1/2} Z \in J(\mathfrak{z})\right) d\mathfrak{z}.\end{aligned}$$

Recall $x^+ = \max(0, x)$ and $x^- = \max(0, -x)$ denote the positive and negative part of $x \in \mathbb{R}$. With $\tilde{Z} = M(\mathrm{y})^{1/2} Z = (\tilde{Z}_1, \tilde{Z}_2)$, we get:

$$\int_{[0,+\infty)^2} \mathbb{P}\left(M(\mathrm{y})^{1/2} Z \in J(\mathfrak{z})\right) d\mathfrak{z} = \mathbb{E}\left[\tilde{Z}_1^- \tilde{Z}_2^-\right]$$

and thus

$$\lim_{n\to\infty} \Delta_n^{(1)} = H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^- \tilde{Z}_2^-\right].$$

Similarly, we obtain:

$$\begin{aligned}\lim_{n\to\infty} \Delta_n^{(2)} &= -H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^- \tilde{Z}_2^+\right] \\ \lim_{n\to\infty} \Delta_n^{(3)} &= -H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^+ \tilde{Z}_2^-\right] \\ \lim_{n\to\infty} \Delta_n^{(4)} &= H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1^+ \tilde{Z}_2^+\right].\end{aligned}$$

Using the definition of $\Sigma_2(\mathrm{y})$ and $M(\mathrm{y})$, notice that $D'(y_1)D'(y_2)\Sigma_2(\mathrm{y})$ is the covariance of $\tilde{Z}_1$ and $\tilde{Z}_2$. Thus, we obtain:

$$\lim_{n\to\infty}\sum_{i=1}^{4}\Delta_n^{(i)} = H(\mathrm{y})\mathbb{E}\left[\left(\tilde{Z}_1^+ - \tilde{Z}_1^-\right)\left(\tilde{Z}_2^+ - \tilde{Z}_2^-\right)\right] = H(\mathrm{y})\mathbb{E}\left[\tilde{Z}_1\tilde{Z}_2\right] = G(\mathrm{y})\Sigma_2(\mathrm{y}). \quad (4.86)$$

Conclusion

Use (4.81), (4.82), (4.84) and (4.86) to get the result. □

The next proposition, the main result of this section, is a consequence of Lemma 4.29.

Proposition 4.30. *Assume that W satisfies condition (4.62). For all $y_1, y_2 \in (0,1)$, we have with $d_1 = D(y_1)$ and $d_2 = D(y_2)$:*

$$\lim_{n\to\infty} n\,\mathbb{E}\left[\left(\mathbf{1}_{\left\{D_1^{(n+1)}\le d_1\right\}} - \mathbf{1}_{\{X_1\le y_1\}}\right)\left(\mathbf{1}_{\left\{D_2^{(n+1)}\le d_2\right\}} - \mathbf{1}_{\{X_2\le y_2\}}\right)\right] = \Sigma_2(y_1,y_2).$$

Proof. Using the comment before Proposition 4.28, we get:

$$\mathbb{E}\left[\prod_{i\in\{1,2\}}\left(\mathbf{1}_{\{D_i^{(n+1)}\le d_i\}} - \mathbf{1}_{\{X_i\le y_i\}}\right)\right] = \mathbb{E}\left[W(X_{[2]})\Psi_{n,-1}(X_{[2]})\right] + \mathbb{E}\left[(1-W(X_{[2]}))\Psi_{n,0}(X_{[2]})\right].$$

We apply Lemma 4.29 twice with $G = W$ and $G = 1 - W$ to get the result. □

4.9 Proof of Theorem 4.23

Recall the definitions of Π_{n+1} and $c_n(y)$ given in (4.61) and (4.67). We define the normalized and centered random process $\hat{\Pi}_{n+1} = (\hat{\Pi}_{n+1}(y) : y \in (0,1))$ by:

$$\hat{\Pi}_{n+1}(y) = \sqrt{n+1}\left[\Pi_{n+1}(y) - c_n(y)\right]. \quad (4.87)$$

Let $U_{n+1} = (U_{n+1}(y) : y \in (0,1))$ be the Hàjek projection of $\hat{\Pi}_{n+1}$:

$$U_{n+1}(y) = \sum_{i=1}^{n+1}\mathbb{E}\left[\hat{\Pi}_{n+1}(y)\middle| X_i\right]. \quad (4.88)$$

Recall Σ defined in Remark 4.24.

Lemma 4.31. *For all $y, z \in (0,1)$, we have:*

$$\lim_{n\to\infty}\mathbb{E}\left[U_{n+1}(y)U_{n+1}(z)\right] = \Sigma(y,z).$$

Proof. Recall (4.71) and (4.72). With $d = D(y)$, we notice that for $y \in (0,1)$:

$$\mathbb{P}\left(D_i^{(n+1)} \le d\middle| X_j\right) - c_n(y) = \begin{cases} \frac{1}{n}H_n(y,X_j) & \text{if } i \ne j, \\ H_n^\star(y,X_j) & \text{if } i = j. \end{cases}$$

We have:

$$\begin{aligned} U_{n+1}(y) &= (n+1)^{-\frac{1}{2}}\sum_{i=1}^{n+1}\sum_{j=1}^{n+1}\left[\mathbb{P}\left(D_i^{(n+1)} \le d\middle| X_j\right) - c_n(y)\right] \\ &= (n+1)^{-\frac{1}{2}}\sum_{j=1}^{n+1}\left[H_n^\star(y,X_j) + H_n(y,X_j)\right]. \end{aligned} \quad (4.89)$$

Let $y, z \in (0,1)$. Since H_n and $H_n^\star$ are centered and $(X_i : i \in \mathbb{N}^*)$ are independent, using (4.89), we obtain that:

$$\mathbb{E}\left[U_{n+1}(y)U_{n+1}(z)\right] = \mathbb{E}\left[H_n^\star(y,X_1)H_n^\star(z,X_1)\right] + \mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(z,X_1)H_n(y,X_1)\right].$$

Recall $\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_3$ defined in Remark 4.24.

Study of $\mathbb{E}\left[H_n^\star(y,X_1)H_n^\star(z,X_1)\right]$

By Proposition 4.27, see (4.74), and by dominated convergence, we get that:

$$\lim_{n\to\infty} \mathbb{E}\left[H_n^\star(y,X_1)H_n^\star(z,X_1)\right] = \mathbb{E}\left[\left(\mathbf{1}_{\{X_1\le y\}} - y\right)\left(\mathbf{1}_{\{X_2\le z\}} - z\right)\right] = \Sigma_1(y,z). \tag{4.90}$$

Study of $\mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right]$

By Proposition 4.27, we have:

$$\begin{aligned}
&\mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right] \\
&\quad= \mathbb{E}\left[\left(\frac{D(y)-W(y,X_1)}{D'(y)} + R_n^{4.27}(y,X_1)\right)\left(\frac{D(z)-W(z,X_1)}{D'(z)} + R_n^{4.27}(z,X_1)\right)\right] \\
&\quad= \frac{1}{D'(y)}\frac{1}{D'(z)}\mathbb{E}\left[(D(y)-W(y,X_1))(D(z)-W(z,X_1))\right] + R_n^{(1)} \\
&\quad= \Sigma_2(y_1,y_2) + R_n^{(1)}
\end{aligned}$$

where, because of (4.73), $\left|R_n^{(1)}\right| \le Cn^{-\frac{1}{4}}$ for some finite constant C. We obtain that

$$\lim_{n\to\infty} \mathbb{E}\left[H_n(y,X_1)H_n(z,X_1)\right] = \Sigma_2(y,z). \tag{4.91}$$

Study of $\mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(z,X_1)H_n(y,X_1)\right]$

By Proposition 4.27, we have that:

$$\mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] = \mathbb{E}\left[H_n^\star(y,X_1)\frac{1}{D'(z)}(D(z)-W(z,X_1)) + H_n^\star(y,X_1)R_n^{4.27}(z,X_1)\right].$$

Thanks to (4.73) and (4.74), we have $|H_n^\star(y,X_1)| \le 1$ and $\mathbb{E}\left[\left|R_n^{4.27}(z,X_1)\right|\right] = O\left(n^{-1/4}\right)$. We deduce from Proposition 4.27 and dominated convergence, that:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] &= \mathbb{E}\left[\left(\mathbf{1}_{\{X_1\le y\}} - y\right)\frac{1}{D'(z)}(D(z)-W(z,X_1))\right] \\
&= \frac{1}{D'(z)}\left(yD(z) - \int_0^y W(z,x)dx\right).
\end{aligned}$$

By symmetry, we finally obtain that

$$\lim_{n\to\infty} \mathbb{E}\left[H_n^\star(y,X_1)H_n(z,X_1)\right] + \mathbb{E}\left[H_n^\star(z,X_1)H_n(y,X_1)\right] = \Sigma_3(y,z). \tag{4.92}$$

Conclusion

Combining (4.90), (4.91) and (4.92), we get that

$$\lim_{n\to\infty} \mathbb{E}\left[U_{n+1}(y)U_{n+1}(z)\right] = \Sigma(y,z).$$

□

Lemma 4.32. *We have the following convergence of finite-dimensional distributions:*

$$(U_{n+1}(y) : y \in (0,1)) \xrightarrow[n\to+\infty]{(fdd)} \chi,$$

where $\chi = (\chi(y) : y \in (0,1))$ *is a centered Gaussian process with covariance function* Σ *given in Remark 4.24.*

Proof. Let $k \in \mathbb{N}^*$ and $(y_1, \dots, y_k) \in (0,1)^k$. We define the random vector $U_{n+1}^{(k)} = (U_{n+1}(y_i) : i \in [k])$. For all $y \in (0,1)$ and $j \in [n+1]$, we set $g_n(y, X_j) = [H_n^\star(y, X_j) + H_n(y, X_j)]$. Using (4.89), we have:

$$U_{n+1}^{(k)} = (n+1)^{-\frac{1}{2}} \sum_{j=1}^{n+1} Z_j^{(n+1)},$$

where $Z_j^{(n+1)} = (g_n(y_i, X_j) : i \in [k])$. Notice $(Z_j^{(n+1)} : j \in [n+1])$ is a sequence of independent, uniformly bounded (see Proposition 4.27) and identically distributed random vectors with mean zero and common positive-definite covariance matrix $V_{n+1} = \mathrm{Cov}\left(Z_1^{(n+1)}\right)$. According to Lemma 4.31, we have that $\lim_{n\to\infty} V_{n+1} = \Sigma^{(k)}$, with $\Sigma^{(k)} = (\Sigma(y_i, y_j) : i, j \in [k])$. The multidimensional Lindeberg-Feller condition is trivially satisfied as $(Z_j^{(n+1)} : j \in [n+1])$ are bounded (uniformly in n) with the same distribution. We deduce from the multidimensional central limit theorem for triangular arrays of random variables, see [22] Corollary 18.2, that $(U_{n+1}^{(k)} : n \geq 0)$ converges in distribution towards the Gaussian random vector with distribution $\mathcal{N}(0, \Sigma^{(k)})$. This gives the result. □

Recall $\hat{\Pi}_n(y)$ defined in (4.87). In view of Lemma 4.32 and since $c_n(y) = y + O(1/n)$, in order to prove Theorem 4.23, it is enough to prove that for all $y \in (0,1)$:

$$\hat{\Pi}_{n+1}(y) - U_{n+1}(y) \xrightarrow[n\to+\infty]{L^2} 0. \tag{4.93}$$

Because $\hat{\Pi}_{n+1}$ and U_{n+1} are centered, we deduce from (4.88) that:

$$\mathbb{E}\left[\left(\hat{\Pi}_{n+1}(y) - U_{n+1}(y)\right)^2\right] = \mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] - \mathbb{E}\left[U_{n+1}(y)^2\right].$$

By Lemma 4.31, we have $\mathbb{E}\left[U_{n+1}(y)^2\right] \underset{n\to\infty}{\longrightarrow} \Sigma(y,y)$. So we deduce that the proof of Theorem 4.23 is a complete as soon as the next lemma is proved.

Lemma 4.33. *For all* $y \in (0,1)$*, we have*

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] = \Sigma(y,y).$$

Proof. Let $y \in (0,1)$ and $d = D(y)$. We have

$$\begin{aligned}
\mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] &= \frac{1}{n+1} \sum_{i,j=1}^{n+1} \mathbb{E}\left[\left(\mathbf{1}_{\left\{D_i^{(n+1)} \leq d\right\}} - c_n(y)\right)\left(\mathbf{1}_{\left\{D_j^{(n+1)} \leq d\right\}} - c_n(y)\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}}\right] - c_n(y)^2 + n\left\{\mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}} \mathbf{1}_{\left\{D_2^{(n+1)} \leq d\right\}}\right] - c_n(y)^2\right\} \\
&= c_n(y) - (n+1)c_n(y)^2 + n\,\mathbb{E}\left[\mathbf{1}_{\left\{D_1^{(n+1)} \leq d\right\}} \mathbf{1}_{\left\{D_2^{(n+1)} \leq d\right\}}\right].
\end{aligned}$$

So we get that

$$\mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] = B_n^{(1)} + B_n^{(2)} + B_n^{(3)} + B_n^{(4)},$$

where

$$\begin{aligned}
B_n^{(1)} &= c_n(y) - c_n(y)^2,\\
B_n^{(2)} &= -n(c_n(y) - y)^2,\\
B_n^{(3)} &= n\,\mathbb{E}\left[\left(\mathbf{1}_{\left\{D_1^{(n+1)}\le d\right\}} - \mathbf{1}_{\{X_1\le y\}}\right)\left(\mathbf{1}_{\left\{D_2^{(n+1)}\le d\right\}} - \mathbf{1}_{\{X_2\le y\}}\right)\right],\\
B_n^{(4)} &= 2n\,\mathbb{E}\left[\mathbf{1}_{\{X_1\le y\}}\left(\mathbf{1}_{\left\{D_2^{(n+1)}\le d\right\}} - c_n(y)\right)\right].
\end{aligned}$$

By Equation (4.70), we get $\lim_{n\to\infty} B_n^{(1)} = \Sigma_1(y,y)$ and $\lim_{n\to\infty} B_n^{(2)} = 0$. By Proposition 4.30, we get $\lim_{n\to\infty} B_n^{(3)} = \Sigma_2(y,y)$. Using (4.71), we get $B_n^{(4)} = 2\,\mathbb{E}\left[\mathbf{1}_{\{X_1\le y\}} H_n(y, X_1)\right]$. By Proposition 4.27 and dominated convergence, we get that:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}_{\{X_1\le y\}} H_n(y,X_1)\right] &= \mathbb{E}\left[\mathbf{1}_{\{X_1\le y\}}\left(\frac{1}{D'(y)}\left(D(y) - W(y,X_1)\right) + R_n^{4.27}(y,X_1)\right)\right]\\
&\xrightarrow[n\to\infty]{} \frac{1}{D'(y)}\left(yD(y) - \int_0^y W(y,x)dx\right).
\end{aligned}$$

This gives $\lim_{n\to\infty} B_n^{(4)} = \Sigma_3(y,y)$. Then, we get that $\lim_{n\to\infty} \mathbb{E}\left[\hat{\Pi}_{n+1}(y)^2\right] = \Sigma(y,y)$. □

✂

Index of notation

$|A|$ cardinality of set A

$[n] = \{1,\dots,n\}$

$|\beta|$ length of $[n]$-word β

$\mathcal{M}_n$ set of $[n]$-words with all characters distinct

$\mathcal{S}_{n,p} = \{\beta \in \mathcal{M}_n : |\beta| = p\}$

$|\mathcal{S}_{n,p}| = A_n^p = n!/(n-p)!$

$\beta_\ell = \beta_{\ell_1}\dots\beta_{\ell_k}$ for $\ell \in \mathcal{S}_{p,k}$ and $\beta \in \mathcal{S}_{n,p}$

$\mathcal{S}_{n,k}^{\ell,\alpha} = \{\beta \in \mathcal{S}_{n,p} : \beta_\ell = \alpha\}$ for $\alpha \in \mathcal{S}_{n,p}$

$|\mathcal{S}_{n,k}^{\ell,\alpha}| = A_{n-k}^{p-k} = (n-k)!/(n-p)!$

$\mathcal{F}$ set of simple finite graphs

F a simple finite graph (and a finite sequence of simple graphs in Sections 3 to 6 satisfying condition (4.38))

$E(F)$ set of edges of F

$V(F)$ set of vertices of F

$v(F) = |V(F)|$ number of vertices of F

$G_n = G_n(W)$ W-random graph with n vertices associated to the sequence $X = (X_k, k \in \mathbb{N}^*)$ of i.i.d. uniform random variables on $[0,1]$

$t(F,G)$ density of hom. from F to G

$t_{\mathrm{inj}}(F,G)$ density of injective hom.

$t_{\mathrm{ind}}(F,G)$ density of embeddings

$Y^\beta(F,G) = \prod_{\{i,j\}\in E(F)} \mathbf{1}_{\{\{\beta_i,\beta_j\}\in E(G)\}}$

$t_{\mathrm{inj}}(F,G) = |\mathcal{S}_{n,p}|^{-1} \sum_{\beta\in\mathcal{S}_{n,p}} Y^\beta(F,G)$

$\boxed{\ell \in \mathcal{M}_p \text{ and } \alpha \in \mathcal{S}_{p,k} \text{ with } k = |\ell|}$

$t_{\text{inj}}(F^\ell, G^\alpha)$ density of injective hom. such that the labelled vertices ℓ of F, with $V(F) = [p]$, are sent on the labelled vertices α of G, with $V(G) = [n]$

$\mathcal{R}_\ell(F)$ sub-graph of the labeled vertices ℓ of F

$Y^\beta(F^\ell, G^\alpha) = Y^\beta(F, G)$ for $\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}$

$\hat{Y}^\alpha(F^\ell, G^\alpha) = \prod_{\{i,j\} \in E(\mathcal{R}_\ell(F))} \mathbf{1}_{\{\{\alpha_i, \alpha_j\} \in E(G)\}}$

$Y^\beta(F^\ell, G^\alpha) = \hat{Y}^\alpha(F^\ell, G^\alpha)\, \tilde{Y}^\beta(F^\ell, G^\alpha)$ i.e.:

$\boxed{Y^\beta = \hat{Y}^\alpha\, \tilde{Y}^\beta}$

$\tilde{t}_{\text{inj}}(F^\ell, G^\alpha) = |\mathcal{S}_{n,p}^{\ell,\alpha}|^{-1} \sum_{\beta \in \mathcal{S}_{n,p}^{\ell,\alpha}} \tilde{Y}^\beta$

$\boxed{t_{\text{inj}}(F^\ell, G^\alpha) = \hat{Y}^\alpha\ \tilde{t}_{\text{inj}}(F^\ell, G^\alpha)}$

$t(F, W)$ hom. densities for graphon W

$t_{\text{ind}}(F, W)$ density of embeddings

$X_\alpha = (X_{\alpha_1}, \dots, X_{\alpha_k})$ and simil. for X_β

$Z^\beta = Z(X_\beta) = \mathbb{E}[Y^\beta(F^\ell, G_n^\alpha)|X]$

$\tilde{Z}^\beta = \mathbb{E}[\tilde{Y}^\beta(F^\ell, G_n^\alpha)|X]$

$t_x = t_x(F^\ell, W) = \mathbb{E}[Z^\beta | X_\alpha = x]$ and

$t_x = \mathbb{E}[t_{\text{inj}}(F^\ell, G_n^\alpha)|X_\alpha = x]$

$\tilde{t}_x = \tilde{t}_x(F^\ell, W) = \mathbb{E}[\tilde{Z}^\beta | X_\alpha = x]$ and

$\tilde{t}_x = \mathbb{E}[\tilde{t}_{\text{inj}}(F^\ell, G_n^\alpha)|X_\alpha = x]$

$\hat{t}_x = \hat{t}_x(F^\ell, W) = \mathbb{E}[\hat{Y}^\alpha(F^\ell, G_n^\alpha)|X_\alpha = x]$

$\boxed{t_x = \hat{t}_x\, \tilde{t}_x}$ for $x \in [0,1]^k$

$t(F^\ell, W) = \int_{[0,1]^k} t_x\, dx = \mathbb{E}[t_{\text{inj}}(F, G_n)]$

$\hat{t}(F^\ell, W) = \int_{[0,1]^k} \hat{t}_x\, dx = \mathbb{E}[t_{\text{inj}}(\mathcal{R}_\ell(F), G_n)]$ and $\hat{t}(F^\ell, W) = t(\mathcal{R}_\ell(F), W)$

$\Gamma_n^{F,\ell}$ random probability measure:

$$\Gamma_n^{F,\ell}(g) = |\mathcal{S}_{n,k}|^{-1} \sum_{\alpha \in \mathcal{S}_{n,k}} g\left(t_{\text{inj}}(F^\ell, G_n^\alpha)\right)$$

$\Gamma^{F,\ell}(dx) = \mathbb{E}\left[\hat{\Gamma}_n^{F,\ell}(dx)\right]$

$\sigma^{F,\ell}(g)^2$ asymptotic variance of

$$\sqrt{n}\left(\Gamma_n^{F,\ell}(g) - \Gamma^{F,\ell}(g)\right)$$

$nD_i^{(n)}$ degree of i in G_n

Π_n empirical CDF of the degrees of G_n

$D(x) = \int_{[0,1]} W(x,y)dy$ degree funct. of W

$\mathcal{H}_{n,d,\delta}(\mathfrak{p}) = \mathbb{P}(X \le nd + \delta)$ for $X \sim \mathcal{B}(n, \mathfrak{p})$

$\sigma_{(x)}^2 = x(1-x)$

$S(x) = \lceil x \rceil - x - \frac{1}{2}$

Φ the CDF of $\mathcal{N}(0,1)$

φ probability distribution density of $\mathcal{N}(01)$

$\boxed{d = D(y)}$

$\boxed{c_n(y) = \mathbb{P}(D_1^{(n+1)} \le d)}$

$H_n^\star(y,u) = \mathbb{P}(D_1^{(n+1)} \le d \,|\, X_1 = u) - c_n(y)$

$\frac{H_n(y,u)}{n} = \mathbb{P}(D_1^{(n+1)} \le d \,|\, X_2 = u) - c_n(y)$

4.10 Appendix A: Preliminary results for the CDF of binomial distributions

In this section, we study uniform asymptotics for the CDF of binomial distributions. Let $n \in \mathbb{N}^*$, $d \in [0,1]$, $\delta \in \mathbb{R}$ and $\mathfrak{p} \in (0,1)$. We consider the CDF:

$$\mathcal{H}_{n,d,\delta}(\mathfrak{p}) = \mathbb{P}\left(X \leq nd + \delta\right), \tag{4.94}$$

where X a binomial random variable with parameters $(n, \mathfrak{p})$. We denote by Φ the cumulative distribution function of the standard Gaussian distribution and by φ the probability distribution density of the standard Gaussian distribution. We recall (4.68) and (4.69): $\sigma^2_{(x)} = x(1-x)$ for $x \in [0,1]$, and $S(x) = \lceil x \rceil - x - \frac{1}{2}$ for $x \in \mathbb{R}$.

We recall a result from [150], see also [186], Chapter VII: for all $x \in \mathbb{R}$, for all $\mathfrak{p} \in (0,1)$ and $n \in \mathbb{N}^*$ such that $n\sigma^2_{(\mathfrak{p})} \geq 25$, we have:

$$\mathbb{P}\left(X \leq n\mathfrak{p} + \sqrt{n}\,\sigma_{(\mathfrak{p})}x\right) = \Phi(x) + \frac{1}{\sqrt{n}}\mathcal{Q}(\mathfrak{p},x) + \frac{1}{\sqrt{n}\,\sigma_{(\mathfrak{p})}}S(n\mathfrak{p} + x\sqrt{n}\,\sigma_{(\mathfrak{p})})\varphi(x) + U_n(\mathfrak{p},x), \tag{4.95}$$

where

$$\mathcal{Q}(\mathfrak{p},x) = \frac{2\mathfrak{p}-1}{6\sigma_{(\mathfrak{p})}}\varphi''(x) = \frac{2\mathfrak{p}-1}{6\sigma_{(\mathfrak{p})}}(x^2-1)\varphi(x),$$

and

$$|U_n(\mathfrak{p},x)| \leq \frac{0.2 + 0.3\,|2\mathfrak{p}-1|}{n\sigma^2_{(\mathfrak{p})}} + \exp\left(-\frac{3\sqrt{n}\,\sigma_{(\mathfrak{p})}}{2}\right). \tag{4.96}$$

We use this result to give an approximation of $\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right)$.

Proposition 4.34. *Let $\varepsilon_0 \in (0, \frac{1}{2})$ and $K_0 = [\varepsilon_0, 1-\varepsilon_0]$. Let $\alpha > 0$. There exists a positive constant C such that for all $n \geq 2$, $s \in [-\alpha\sqrt{\log(n)}, \alpha\sqrt{\log(n)}]$, $\delta \in [-1,1]$ and $d \in K_0$ such that $d + \frac{s}{\sqrt{n}} \in K_0$, we have:*

$$\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right) = \Phi(y_s) + \frac{1}{\sqrt{n}}\frac{\varphi(y_s)}{\sigma_{(d)}}\pi(s,n,d,\delta) + R^{4.34}(s,n,d,\delta),$$

where

$$y_s = \frac{-s}{\sigma_{(d)}} \quad \text{and} \quad \pi(s,n,d,\delta) = \frac{1-2d}{6}(1+2y_s^2) + S(nd+\delta) + \delta \tag{4.97}$$

and

$$\left|R^{4.34}(s,n,d,\delta)\right| \leq C\frac{\log(n)^2}{n}.$$

Proof. In what follows, C denotes a positive constant which depends on ε_0 (but not on $n \geq 2$, $s \in [-\alpha\sqrt{\log(n)}, \alpha\sqrt{\log(n)}]$, $\delta \in [-1,1]$ and $d \in K_0$ such that $d + \frac{s}{\sqrt{n}} \in K_0$) and which may change from line to line. We will also use, without recalling it, that $\sigma_{(.)}$ is uniformly bounded away from 0 on K_0.

For all $\theta \in (0,1]$ such that $d + s\theta \in (0,1)$, we set

$$x_s(\theta) = \frac{-s + \delta\theta}{\sigma_{(d+s\theta)}}. \tag{4.98}$$

Let $\mathfrak{p} = d + \frac{s}{\sqrt{n}}$ and X be a binomial random variable with parameters $(n, \mathfrak{p})$. Because $nd + \delta = n\mathfrak{p} + \sqrt{n}\sigma_{(\mathfrak{p})}\,x_s\left(\frac{1}{\sqrt{n}}\right)$, we can write

$$\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right) = \mathbb{P}\left(X \leq n\mathfrak{p} + \sqrt{n}\,\sigma_{(\mathfrak{p})}\,x_s\left(\frac{1}{\sqrt{n}}\right)\right). \tag{4.99}$$

Recall S is defined in (4.69). Using (4.95), we get that:

$$\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right)=\Phi\left(x_s\left(\frac{1}{\sqrt{n}}\right)\right)+\frac{1}{\sqrt{n}}\,Q^{(1)}_{d,\delta}\left(s,\frac{1}{\sqrt{n}}\right)+\frac{1}{\sqrt{n}}\,Q^{(2)}_{d,\delta}\left(s,\frac{1}{\sqrt{n}}\right)+U_n\left(d+\frac{s}{\sqrt{n}},x_s\left(\frac{1}{\sqrt{n}}\right)\right) \tag{4.100}$$

where for $\theta\in(0,1]$ such that $d+s\theta\in(0,1)$,

$$Q^{(1)}_{d,\delta}(s,\theta)=\frac{2(d+s\theta)-1}{6\sigma_{(d+s\theta)}}\left(x_s(\theta)^2-1\right)\varphi(x_s(\theta)),$$
$$Q^{(2)}_{d,\delta}(s,\theta)=\frac{1}{\sigma_{(d+s\theta)}}S\left(d\theta^{-2}+\delta\right)\varphi(x_s(\theta)).$$

Study of the first term on the right-hand side of (4.100)

Let $\theta\in(0,1/\sqrt{2}]$, and notice that $|\log(\theta)|\geq\log(\sqrt{2})>0$. Recall the definition of $x_s(\theta)$ given by (4.98). By simple computations, we get that for all $0<\theta\leq 1/\sqrt{2}$, $|s|\leq\alpha\sqrt{2|\log(\theta)|}$, $|\delta|\leq 1$, and $d\in K_0$ such that $d+s\theta\in K_0$,

$$|x_s(\theta)|\leq C\,|\log(\theta)|^{\frac{1}{2}},\quad |x_s'(\theta)|\leq C\,|\log(\theta)|\quad\text{and}\quad |x_s''(\theta)|\leq C\,|\log(\theta)|^{\frac{3}{2}}. \tag{4.101}$$

We define the function $\Psi_s(\theta)=\Phi(x_s(\theta))$. Applying Taylor's theorem with the Lagrange form of the remainder for Ψ at $\theta=0$, we have:

$$\Psi_s(\theta)=\Psi_s(0)+\theta\Psi_s'(0)+R_s^{(1)}(\theta),$$

where $R_s^{(1)}(\theta)=\int_0^\theta\Psi_s''(t)(\theta-t)dt$. Recall the definition of $y_s=x_s(0)$ given in (4.97). Elementary calculus gives:

$$\begin{aligned}\Phi\left(x_s(\theta)\right)&=\Phi(x_s(0))+\theta\,x_s'(0)\varphi(x_s(0))+R_s^{(1)}(\theta)\\&=\Phi\left(y_s\right)+\theta\left[\frac{(1-2d)}{2\sigma_{(d)}}y_s^2+\frac{\delta}{\sigma_{(d)}}\right]\varphi\left(y_s\right)+R_s^{(1)}(\theta),\end{aligned} \tag{4.102}$$

where $R_s^{(1)}(\theta)=\int_0^\theta\Big(x_s''(t)-x_s'(t)^2x_s(t)\Big)\varphi(x_s(t))\,(\theta-t)\,dt$. Using (4.101) and that $t\varphi(t)$ is bounded, we have:

$$\left|R_s^{(1)}(\theta)\right|\leq C\theta^2(|\log(\theta)|^{\frac{3}{2}}+|\log(\theta)|^2)\leq C\theta^2|\log(\theta)|^2.$$

Study of the second term on the right-hand side of (4.100)

We have $Q^{(1)}_{d,\delta}(s,\theta)=G_s(\theta)H(x_s(\theta))$ where

$$G_s(\theta)=\frac{2(d+s\theta)-1}{6\sigma_{(d+s\theta)}}\quad\text{and}\quad H(x)=\left(x^2-1\right)\varphi(x).$$

For the first term, we have

$$G_s(\theta)=G_s(0)+R_s^{(2)}(\theta)=\frac{2d-1}{6\sigma_{(d)}}+R_s^{(2)}(\theta),$$

where $R_s^{(2)}(\theta)=\int_0^\theta G_s'(t)dt$. We compute that:

$$G_s'(t)=s\left[\frac{1}{3\sigma_{(d+st)}}+\frac{[2(d+st)-1]^2}{12\sigma^3_{(d+st)}}\right].$$

We obtain that $\left|R_s^{(2)}(\theta)\right| \leq C\theta|s| \leq C\theta|\log(\theta)|^{\frac{1}{2}}$. For the second term, we have

$$H(x_s(\theta)) = H(x_s(0)) + R_s^{(3)}(\theta) = \left(y_s^2 - 1\right)\varphi\left(y_s\right) + R_s^{(3)}(\theta),$$

where $R_s^{(3)}(\theta) = \int_0^\theta x_s'(t)H'(x_s(t))dt = \int_0^\theta x_s'(t)\left[-x_s(t)^3 + 3x_s(t)\right]\varphi(x_s(t))dt$. Using (4.101) and that $(|t|^3+t)\varphi(t)$ is bounded, we get that $\left|R_s^{(3)}(\theta)\right| \leq C\theta|\log(\theta)|$. Finally, we obtain that

$$Q_{d,\delta}^{(1)}(s,\theta) = \frac{2d-1}{6\sigma_{(d)}}\left(y_s^2 - 1\right)\varphi\left(y_s\right) + R_s^{(4)}(\theta) \tag{4.103}$$

with $\left|R_s^{(4)}(\theta)\right| \leq C\theta|\log(\theta)|$.

Study of the last term on the right-hand side of (4.100)

We have

$$Q_{d,\delta}^{(2)}(s,\theta) = F_s(\theta)S\left(\frac{d}{\theta^2} + \delta\right)\varphi(x_s(\theta)) \quad \text{with} \quad F_s(\theta) = \frac{1}{\sigma_{(d+s\theta)}}.$$

For the first term on the right-hand side, we have

$$F_s(\theta) = F_s(0) + R_s^{(5)}(\theta) = \frac{1}{\sigma_{(d)}} + R_s^{(5)}(\theta),$$

where $R_s^{(5)}(\theta) = \int_0^\theta F_s'(t)dt = \int_0^\theta \frac{s(2(d+st)-1)}{2\sigma_{(d+st)}^3}dt$. We get that $\left|R_s^{(5)}(\theta)\right| \leq C\theta|s| \leq C\theta|\log(\theta)|^{\frac{1}{2}}$. For the last term on the right-hand side, we have:

$$\varphi(x_s(\theta)) = \varphi(x_s(0)) + R_s^{(6)}(\theta) = \varphi\left(y_s\right) + R_s^{(6)}(\theta),$$

where $R_s^{(6)}(\theta) = \int_0^\theta x_s'(t)\varphi'(x_s(t))dt = -\int_0^\theta x_s(t)x_s'(t)\varphi(x_s(t))dt$. So, using (4.101) and that $t\varphi(t)$ is bounded, we get that $\left|R_s^{(6)}(\theta)\right| \leq C\theta|\log(\theta)|$. Finally, we obtain that

$$Q_{d,\delta}^{(2)}(s,\theta) = \frac{1}{\sigma_{(d)}}S\left(\frac{d}{\theta^2} + \delta\right)\varphi\left(y_s\right) + R_s^{(7)}(\theta), \tag{4.104}$$

where $\left|R_s^{(7)}(\theta)\right| \leq C\theta\left|\log(\theta)\right|$, since S is bounded.

Conclusion

We deduce from (4.102), (4.103) and (4.104) that

$$\Phi\left(x_s(\theta)\right) + \theta\, Q_{d,\delta}^{(1)}(s,\theta) + \theta\, Q_{d,\delta}^{(2)}(s,\theta) = \Phi(y_s) + \theta\frac{\varphi(y_s)}{\sigma_{(d)}}\pi\left(s, \frac{1}{\theta^2}, d, \delta\right) + R_s^{(8)}(\theta),$$

where $\left|R_s^{(8)}(\theta)\right| \leq C\theta^2|\log(\theta)|^2$. We get the result by taking $\theta = 1/\sqrt{n}$ and using (4.100) and the obvious bound on U_n given by (4.96) so that $|U_n| \leq C/n$. □

We state a Lemma which will be useful for the proof of Corollary 4.36.

Lemma 4.35. *Let $y \in [0,1]$ and $\alpha > 0$. For all $n \geq 2$, we have with $d = D(y)$, $A = \alpha\sqrt{\log(n)}$ and $y_s = -s/\sigma_{(d)}$,*

$$\Phi\left(-\frac{A}{\sigma_{(d)}}\right) \leq \frac{1}{\alpha\, n^{2\alpha^2}}, \quad \int_A^{+\infty} s\Phi\left(y_s\right)ds \leq \frac{1}{\alpha\, n^{2\alpha^2}} \quad \text{and} \quad \int_A^{+\infty} s^2\varphi(y_s)ds \leq \frac{1}{\alpha\, n^{\alpha^2}}.$$

Proof. For all $t \geq 0$, we have

$$\Phi(-t) = \int_t^{+\infty} s\frac{\varphi(s)}{s}ds \leq \frac{1}{t}\int_t^{+\infty} s\varphi(s)ds = \frac{1}{t}\varphi(t). \tag{4.105}$$

Because $\sigma_{(d)} \leq 1/2$, we get with $t = \frac{A}{\sigma_{(d)}}$ the following rough upper bound:

$$\Phi\left(-\frac{A}{\sigma_{(d)}}\right) \leq \frac{\sigma_{(d)}}{A}\varphi\left(\frac{A}{\sigma_{(d)}}\right) \leq \frac{1}{\alpha\, n^{2\alpha^2}}\,. \tag{4.106}$$

Using again (4.105) and (4.106), we get, for the second inequality that:

$$\int_A^{+\infty} s\Phi\left(y_s\right) ds \leq \sigma_{(d)}\int_A^{+\infty} \varphi(-y_s)ds = \sigma_{(d)}^2\Phi\left(-\frac{A}{\sigma_{(d)}}\right) \leq \frac{1}{\alpha\, n^{2\alpha^2}}\,.$$

For the last inequality, we have:

$$\begin{aligned}\int_A^{+\infty} s^2\varphi(y_s)ds = \frac{2\sigma_{(d)}^2}{\sqrt{2\pi}}\int_A^{+\infty}\frac{s^2}{2\sigma_{(d)}^2}\mathrm{e}^{-\frac{s^2}{2\sigma_{(d)}^2}}ds &\leq \frac{4\sigma_{(d)}^2}{\sqrt{2\pi}}\int_A^{+\infty}\mathrm{e}^{-\frac{s^2}{4\sigma_{(d)}^2}}ds\\ &= 4\sqrt{2}\sigma_{(d)}^3\Phi\left(-\frac{A}{\sqrt{2}\sigma_{(d)}}\right)\\ &\leq \frac{1}{\alpha\, n^{\alpha^2}},\end{aligned}$$

where we used $x\mathrm{e}^{-x} \leq 2\mathrm{e}^{-\frac{x}{2}}$ for the first inequality and an inequality similar to (4.106) with $\sigma_{(d)}$ replaced by $\sqrt{2}\sigma_{(d)}$ for the last one. □

For $f \in \mathcal{C}^2([0,1])$, we set $\|f\|_{3,\infty} = \|f\|_\infty + \|f'\|_\infty + \|f''\|_\infty$.

Lemma 4.36. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$ and $\alpha \geq 1$. There exists a positive constant C such that for all $H \in \mathcal{C}^2([0,1])$, $\delta \in [-1,1]$ and $n \geq 2$ such that $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D\big((0,1)\big)$, with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$, we have:*

$$\begin{aligned}\sqrt{n}\int_{-A}^{A} H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\leq 0\}}\right)ds&\\ = \frac{H'(y)}{D'(y)}\frac{\sigma_{(d)}^2}{2} + H(y)\left(\frac{1-2d}{2} + \delta + S(nd+\delta)\right) + R_n^{4.36}(H),&\end{aligned}$$

where

$$\left|R_n^{4.36}(H)\right| \leq C\,\|H\|_{3,\infty}\, n^{-1/2}\log(n)^3. \tag{4.107}$$

Because of the assumption $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D\big((0,1)\big)$, we need to rule out the cases $y \in \{0,1\}$, so that Lemma 4.36 holds only for $y \in (0,1)$.

Proof. In what follows, C denotes a positive constant which depends on ε_0 and W (but in particular not on $n \geq 2$, $s \in [-\alpha\sqrt{\log(n)}, \alpha\sqrt{\log(n)}]$, $\delta \in [-1,1]$ and $d \in K_0$ such that $d + \frac{s}{\sqrt{n}} \in K_0$) and which may change from lines to lines.

Let $\theta \in (0, 1/\sqrt{2}]$ (we shall take $\theta = 1/\sqrt{n}$ later on) and assume that $|s| \leq \alpha\sqrt{2|\log(\theta)|}$ and $d + s\theta \in K_0$. We set $\Psi(\theta) = H\left(D^{-1}(d+s\theta)\right)$. Notice that $\Psi'(\theta) = \frac{s}{D'\circ D^{-1}(d+s\theta)}H'(D^{-1}(d+s\theta))$. By Taylor's theorem with the Lagrange form of the remainder we have:

$$\Psi(\theta) = \Psi(0) + \theta\Psi'(0) + R_s^{(1)}(\theta) = H(y) + \theta\frac{s}{D'(y)}H'(y) + R_s^{(1)}(\theta) \tag{4.108}$$

where $R_s^{(1)}(\theta) = \int_0^\theta \Psi''(t)(\theta - t)dt$. We have

$$\Psi''(\theta) = s^2 \left[\frac{H''(D^{-1}(d+s\theta))}{(D' \circ D^{-1}(d+s\theta))^2} - \frac{H'(D^{-1}(d+s\theta))(D'' \circ D^{-1}(d+s\theta))}{(D' \circ D^{-1}(d+s\theta))^3} \right].$$

Thus, we get that $\left|R_s^{(1)}(\theta)\right| \leq C\ (\|H'\|_\infty + \|H''\|_\infty)\ s^2\theta^2 \leq C\,\|H\|_{3,\infty}\,\theta^2|\log(\theta)|$. Choosing $\theta = 1/\sqrt{n}$, we deduce from (4.108) that:

$$H\left(D^{-1}\left(d + \frac{s}{\sqrt{n}}\right)\right) = H(y) + \frac{1}{\sqrt{n}}\frac{s}{D'(y)}H'(y) + R_s^{(1)}\left(\frac{1}{\sqrt{n}}\right), \tag{4.109}$$

where $\left|R_s^{(1)}\left(1/\sqrt{n}\right)\right| \leq C\,\|H\|_{3,\infty}\,\log(n)/n$. Recall the definition of y_s and $\pi(s,n,d,\delta)$ given by (4.97). By Proposition 4.34 and equation (4.109), we get that:

$$\begin{aligned}
&\sqrt{n}\,H\left(D^{-1}\left(d + \frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d + \frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\leq 0\}}\right) \\
&= \sqrt{n}\left(H(y) + \frac{1}{\sqrt{n}}\frac{s}{D'(y)}H'(y)\right)\left(\left(\Phi(y_s) - \mathbf{1}_{\{s\leq 0\}}\right) + \frac{1}{\sqrt{n}}\frac{\varphi(y_s)}{\sigma_{(d)}}\pi(s,n,d,\delta)\right) + R_n^{(0)}(s) \\
&= \sqrt{n}\,H(y)\Delta^{(1)}(s) + \frac{H'(y)}{D'(y)}\Delta^{(2)}(s) + \frac{H(y)}{\sigma_{(d)}}\Delta^{(3)}(s) + R_n^{(0)}(s) + \hat{R}_n^{(0)}(s),
\end{aligned} \tag{4.110}$$

where

$$\Delta^{(1)}(s) = \left(\Phi(y_s) - \mathbf{1}_{\{s\leq 0\}}\right), \quad \Delta^{(2)}(s) = s\left(\Phi(y_s) - \mathbf{1}_{\{s\leq 0\}}\right), \quad \Delta^{(3)}(s) = \varphi(y_s)\pi(s,n,d,\delta),$$

$$\begin{aligned}
\left|R_n^{(0)}(s)\right| &\leq \sqrt{n}\|H\|_\infty|R^{4.34}(s,n,d,\delta)| + \sqrt{n}|R_s^{(1)}\left(1/\sqrt{n}\right)| + \sqrt{n}|R^{4.34}(s,n,d,\delta)R_s^{(1)}\left(1/\sqrt{n}\right)| \\
&\leq C\,\|H\|_{3,\infty}\frac{\log(n)^2}{\sqrt{n}}
\end{aligned} \tag{4.111}$$

and

$$\left|\hat{R}_n^{(0)}(s)\right| = \left|\frac{1}{\sqrt{n}}\frac{H'(y)}{\sigma_{(d)}D'(y)}s\varphi(y_s)\pi(s,n,d,\delta)\right| \leq C\,\|H\|_{3,\infty}\frac{\sqrt{\log(n)}}{\sqrt{n}}. \tag{4.112}$$

Study of $\int_{-A}^{A}\Delta^{(1)}(s)ds$

Since $\Delta^{(1)}$ is an odd integrable function on $\mathbb{R}^*$, we get that:

$$\int_{-A}^{A}\Delta^{(1)}(s)ds = 0. \tag{4.113}$$

Study of $\int_{-A}^{A}\Delta^{(2)}(s)ds$

Because $\Delta^{(2)}$ is integrable and $\int_{\mathbb{R}}\Delta^{(2)}(s)ds = \sigma_{(d)}^2/2$, we get that

$$\int_{-A}^{A}\Delta^{(2)}(s)\,ds = \frac{\sigma_{(d)}^2}{2} + R_n^{(2)}, \quad \text{with} \quad R_n^{(2)} = -2\int_A^{+\infty} s\Phi(y_s)ds. \tag{4.114}$$

Using Lemma 4.35, we get that

$$|R_n^{(2)}| \leq C\,n^{-2\alpha^2}. \tag{4.115}$$

Study of $\int_{-A}^{A} \Delta^{(3)}(s)ds$

We have, using (4.97), that:

$$\Delta^{(3)}(s) = \varphi(y_s)\left(\frac{1-2d}{6}(1+2y_s^2) + \delta + S(nd+\delta)\right).$$

By elementary calculus, we have that:

$$\int_{\mathbb{R}} \varphi(y_s)ds = \int_{\mathbb{R}} y_s^2\varphi(y_s)ds = \sigma_{(d)}.$$

We get that:

$$\int_{-A}^{A} \Delta^{(3)}(s)ds = \sigma_{(d)}\left[\frac{1-2d}{2} + \delta + S\left(nd+\delta\right)\right] + R_n^{(3)}, \tag{4.116}$$

where

$$R_n^{(3)} = -2\sigma_{(d)}\left(\frac{1-2d}{6} + \delta + S(nd+\delta)\right)\Phi\left(-\frac{A}{\sigma_{(d)}}\right) - 2\,\frac{1-2d}{3}\int_A^{+\infty} y_s^2\varphi(y_s)ds. \tag{4.117}$$

Using Lemma 4.35 and since $|2d-1| \le 1$, $|\delta| \le 1$ and S is bounded by 1, we have that:

$$|R_n^{(3)}| \le Cn^{-\alpha^2}. \tag{4.118}$$

Conclusion

Using (4.110), (4.113), (4.114), (4.116), we deduce that

$$\sqrt{n}\int_{-A}^{A} H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\le 0\}}\right)ds$$
$$= \frac{H'(y)}{D'(y)}\frac{\sigma_{(d)}^2}{2} + H(y)\left[\frac{1-2d}{2} + \delta + S\left(nd+\delta\right)\right] + R_n^{4.36}(H),$$

where $R_n^{4.36}(H) = \int_{-A}^{A}(R_n^{(0)}(s) + \hat{R}_n^{(0)}(s))\,ds + (H'(y)/D'(y))R_n^{(2)} + (H(y)/\sigma_{(d)})R_n^{(3)}$. Using the upper bounds (4.111) and (4.112) (to be integrated over $[-A,A]$), (4.115) and (4.118) with $\alpha \ge 1$, we get that $|R_n^{4.36}(H)| \le C\,\|H\|_{3,\infty}\log(n)^3/\sqrt{n}$. □

We give a direct application of the previous lemma.

Lemma 4.37. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$ and $\alpha \ge 1$. There exists a positive constant C such that for all $G \in \mathcal{C}^2([0,1])$, $\delta \in [-1,1]$, $n \ge 2$ such that $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D((0,1))$, with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$, we have:*

$$n\int_0^1 G(x)\left(\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x\le y\}}\right)\mathbf{1}_{\{\sqrt{n}|D(x)-d|\le A\}}dx$$
$$= \frac{G'(y)D'(y) - G(y)D''(y)}{D'(y)^3}\frac{\sigma_{(d)}^2}{2} + \frac{G(y)}{D'(y)}\left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.37}(G),$$

where

$$\left|R_n^{4.37}(G)\right| \le C\,\|G\|_{3,\infty}\,n^{-1/2}\log(n)^3.$$

Proof. Let G be a function in $\mathcal{C}^2([0,1])$. Define the function H on $[0,1]$ by $H(z) = \frac{G(z)}{D'(z)}$ for all $z \in [0,1]$. Use the change of variables $s = \sqrt{n}\,(D(x) - d)$ to get that:

$$\int_0^1 G(x)\left(\mathcal{H}_{n,d,\delta}(D(x)) - \mathbf{1}_{\{x\le y\}}\right)\mathbf{1}_{\{\sqrt{n}|D(x)-d|\le A\}}dx$$
$$= \frac{1}{\sqrt{n}}\int_{-A}^{A} H\left(D^{-1}\left(d+\frac{s}{\sqrt{n}}\right)\right)\left(\mathcal{H}_{n,d,\delta}\left(d+\frac{s}{\sqrt{n}}\right) - \mathbf{1}_{\{s\le 0\}}\right)ds.$$

By Lemma 4.36, we obtain that:

$$n\int_0^1 G(x)\left(\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x\le y\}}\right)\mathbf{1}_{\left\{\sqrt{n}|D(x)-d|\le A\right\}}dx$$
$$= \frac{H'(y)}{D'(y)}\frac{\sigma^2_{(d)}}{2} + H(y)\left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.36}(H)$$
$$= \frac{G'(y)D'(y) - G(y)D''(y)}{D'(y)^3}\frac{\sigma^2_{(d)}}{2} + \frac{G(y)}{D'(y)}\left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.36}(G/D').$$

Set $R_n^{4.37}(G) = R_n^{4.36}(G/D')$ and use (4.107) to conclude. □

Lemma 4.38. *Let $y \in (0,1)$ and $\alpha > 0$. For all $u \in (0,1)$, $\delta \in [-1,1]$ and $n \in \mathbb{N}^*$ such that $\sqrt{n}|u-d| \ge A$ with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$, we have*

$$\left|\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\le d\}}\right| \le n^{-\alpha+2}.$$

Proof. Let X be a binomial random variable with parameters (n,u). Assume first that $u \ge d + \frac{A}{\sqrt{n}}$. Let $\lambda \ge 0$. Using Chernov's inequality, we get:

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\le d\}} = \mathbb{P}\left(X \le nd+\delta\right) \le \mathrm{e}^{\lambda(nd+\delta)}\,\mathbb{E}\left[\mathrm{e}^{-\lambda X}\right] = \exp\left[\lambda(nd+\delta) + n\Psi(\lambda)\right], \tag{4.119}$$

with $\Psi(\lambda) = \log(1+u(\mathrm{e}^{-\lambda}-1))$. By Taylor's theorem with the Lagrange form of the remainder, we have

$$\Psi(\lambda) = \Psi(0) + \lambda\Psi'(0) + R(\lambda) = 0 - u\lambda + R(\lambda), \tag{4.120}$$

where $R(\lambda) = \int_0^\lambda (\lambda - t)\Psi''(t)dt$. Because $\Psi''(t) \ge 0$ and $\Psi''(t) = \frac{(1-u)u\mathrm{e}^{-\lambda}}{\left(1+u(\mathrm{e}^{-\lambda}-1)\right)^2} \le \frac{1}{4}$ (applying the following inequality $\frac{xy}{(x+y)^2} \le \frac{1}{4}$ with $x = 1-u$ and $y = u\mathrm{e}^{-\lambda}$), we get that $|R(\lambda)| \le \frac{\lambda^2}{8} \le \lambda^2$. Finally, applying (4.120) with $\lambda = \sqrt{\frac{\log(n)}{n}}$, we get that

$$n\Psi(\lambda) = -u\sqrt{n\log(n)} + R^{(2)}(n), \tag{4.121}$$

with $|R^{(2)}(n)| \le \log(n)$. Using (4.119) and (4.121), we get that

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\le d\}} \le \exp\left[\sqrt{\frac{\log(n)}{n}}(nd+\delta) - u\sqrt{n\log(n)} + R^{(2)}(n)\right]$$
$$= \exp\left[\sqrt{n\log(n)}(d-u) + R^{(3)}(n)\right],$$

where $|R^{(3)}(n)| \le 2\log(n)$, since $|\delta| \le 1$. Because $d - u \le \frac{-A}{\sqrt{n}}$ with $A = \alpha\sqrt{\log(n)}$, we have that

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u\le d\}} \le \mathrm{e}^{-\alpha\log(n) + R^{(3)}(n)} \le \mathrm{e}^{(-\alpha+2)\log(n)} = n^{-\alpha+2}.$$

In the case where $u \leq d - \frac{A}{\sqrt{n}}$, we have that

$$\begin{aligned} 0 \geq \mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u \leq d\}} &= \mathbb{P}\left(X \leq nd + \delta\right) - 1 \\ &\geq -\mathbb{P}\left(X \geq nd + \delta\right) \\ &= -\mathbb{P}\left(n - X \leq n(1-d) - \delta\right). \end{aligned}$$

Since $n - X$ is a binomial random variable with parameters $(n, 1-u)$, using similar argument as in the first part of the proof (with u and X replaced by $1-u$ and $n-X$), we get that, for $u \leq d - \frac{A}{\sqrt{n}}$:

$$\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u \leq d\}} \geq -n^{-\alpha+2}.$$

We deduce that $\left|\mathcal{H}_{n,d,\delta}(u) - \mathbf{1}_{\{u \leq d\}}\right| \leq n^{-\alpha+2}$. □

The following lemma is a direct application of Lemma 4.38 with $u = D(x)$.

Lemma 4.39. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$ and $\alpha \geq 1$. For all $G \in \mathcal{B}([0,1])$, $\delta \in [-1,1]$ and $n \in \mathbb{N}^*$, we have with $d = D(y)$ and $A = \alpha\sqrt{\log(n)}$:*

$$n \int_0^1 G(x) \left|\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x \leq y\}}\right| \mathbf{1}_{\{\sqrt{n}|D(x)-d| \geq A\}} dx = R_n^{4.39}(G),$$

where

$$\left|R_n^{4.39}(G)\right| \leq \|G\|_\infty \, n^{-\alpha+3}.$$

Combining Lemma 4.37 with Lemma 4.39 for $\alpha = 3$, we deduce the following proposition.

Proposition 4.40. *Assume that W satisfies condition (4.62). Let $y \in (0,1)$. There exists a positive constant C such that for all $G \in \mathcal{C}^2([0,1])$, $\delta \in [-1,1]$ and $n \in \mathbb{N}^*$ such that $\left[d \pm \frac{A}{\sqrt{n}}\right] \subset D((0,1))$, with $d = D(y)$ and $A = 4\sqrt{\log(n)}$, we have:*

$$\begin{aligned} &n \int_0^1 G(x) \left(\mathcal{H}_{n,d,\delta}\left(D(x)\right) - \mathbf{1}_{\{x \leq y\}}\right) dx \\ &\qquad = \frac{G'(y)D'(y) - G(y)D''(y)}{D'(y)^3} \frac{\sigma_{(d)}^2}{2} + \frac{G(y)}{D'(y)} \left[\frac{1-2d}{2} + \delta + S(nd+\delta)\right] + R_n^{4.40}(G), \end{aligned}$$

with

$$\left|R_n^{4.40}(G)\right| \leq C \, \|G\|_{3,\infty} \, n^{-\frac{1}{4}}. \tag{4.122}$$

4.11 Appendix B: Proof of Proposition 4.28

We first state a preliminary lemma in Section 4.11.1 and then provide the proof of Proposition 4.28 in Section 4.11.2.

4.11.1 A preliminary result

For $\mathrm{y} = (y_1, y_2) \in [0,1]^2$, let $M(\mathrm{y})$ be the covariance matrix of a pair (Y_1, Y_2) of Bernoulli random variables such that $\mathbb{P}(Y_i = 1) = D(y_i)$ for $i \in \{1,2\}$ and $\mathbb{P}(Y_1 = Y_2 = 1) = \int_{[0,1]} W(y_1, z) W(y_2, z) \, dz$.

Lemma 4.41. *Assume that W satisfies condition (4.62). There exists $\varepsilon' > 0$ such that for all $\mathrm{y} \in [0,1]^2$, we have $\det(M(\mathrm{y})) > \varepsilon'$.*

Proof. Let $\mathbb{M}_2$ be the set of matrices of size 2×2, and $\|\cdot\|_\infty$ be the norm on $\mathbb{M}_2$ defined in (4.127). We consider the closed set on $\mathbb{M}_2$:

$$\mathcal{F} = \mathcal{F}_+ \bigcup \mathcal{F}_- \quad \text{where} \quad \mathcal{F}_\pm = \left\{ r(I_2 \pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}); \, r \in [0, 1/4] \right\}$$

where $I_2 \in \mathbb{M}_2$ is the identity matrix. Notice $\mathcal{F}$ is the set of all covariance matrices of pairs of Bernoulli random variables having determinant equal to 0. Since the determinant is a continuous real-valued function on $\mathbb{M}_2$, to prove Lemma 4.41, it is enough to prove that for all $\mathrm{y} = (y, y') \in [0,1]^2$ and all $M_0 \in \mathcal{F}$:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4. \tag{4.123}$$

We set $p = D(y)$, $p' = D(y')$ and $\alpha = \int_{[0,1]} W(y,z)W(y',z)\,dz$ so that:

$$M(\mathrm{y}) = \begin{pmatrix} p(1-p) & \alpha - pp' \\ \alpha - pp' & p'(1-p') \end{pmatrix}.$$

The proof of (4.123) is divided into three cases. Recall that W satisfies condition (4.62). Without loss of generality, we can assume that $p \leq p'$ and thus:

$$\varepsilon \leq p \leq p' \leq 1 - \varepsilon. \tag{4.124}$$

Since $(1 - W(y,z))(1 - W(y',z))$ is non-negative, by integrating with respect to z over $[0,1]$, we get that $\alpha \geq p + p' - 1$. Using that $W \leq 1 - \varepsilon$, we deduce, denoting by $x^+ = \max(x, 0)$ the positive part of $x \in \mathbb{R}$, that:

$$(p + p' - 1)_+ \leq \alpha \leq (1 - \varepsilon)p. \tag{4.125}$$

The case $M_0 \in \mathcal{F}_+$

Recall that $p \leq p'$. If $|r - p(1-p)| \geq \varepsilon^2/4$, then, by considering the first term on the diagonal, we have $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

If $|r - p(1-p)| \leq \varepsilon^2/4$, then, by considering the off-diagonal term, we have:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq |\alpha - pp' - r|.$$

For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, we get, using that $\alpha \leq (1-\varepsilon)p$ and $p \leq p'$:

$$\begin{aligned} \alpha - pp' - r &\leq (1-\varepsilon)p - p^2 - p(1-p) - \delta' \\ &\leq -\varepsilon^2 + \varepsilon^2/4 = -3\varepsilon^2/4. \end{aligned}$$

We deduce that (4.123) holds if $M_0 \in \mathcal{F}_+$.

The case $|1 - p - p'| > \varepsilon/2$ and $M_0 \in \mathcal{F}_-$

If $|r - p(1-p)| \geq \varepsilon^2/4$, then, by considering the first term on the diagonal, we have $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

If $|r - p(1-p)| \leq \varepsilon^2/4$, then, by considering the off-diagonal term, we have:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq |\alpha - pp' + r|.$$

Assume first that $1 - p - p' > \varepsilon/2$. For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, we get, using $\alpha \geq 0$, that:

$$\begin{aligned} \alpha - pp' + r &\geq p(1 - p - p') + \delta' \\ &\geq \varepsilon^2/2 - \varepsilon^2/4 = \varepsilon^2/4. \end{aligned}$$

Assume then that $1-p-p' < -\varepsilon/2$. For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, we get, using the lower bound $\alpha \geq p + p' - 1$ from (4.125), that:

$$\begin{aligned}\alpha - pp' + r &\geq (1-p)(p+p'-1) + \delta' \\ &\geq \varepsilon^2/2 - \varepsilon^2/4 = \varepsilon^2/4.\end{aligned}$$

We get $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

We deduce that (4.123) holds if $|1-p-p'| > \varepsilon/2$ and $M_0 \in \mathcal{F}_-$.

The case $|1-p-p'| \leq \varepsilon/2$ and $M_0 \in \mathcal{F}_-$

Applying Lemma 4.42 below, with $f = W(y,\cdot)$, $g = W(y',\cdot)$ and $\delta = 1-p-p'$, we get that:

$$\alpha \geq (1-\varepsilon)(\varepsilon - \delta). \tag{4.126}$$

If $|r - p(1-p)| \geq \varepsilon^2/4$, then, by considering the first term on the diagonal, we have $\| M(\mathrm{y}) - M_0 \|_\infty \geq \varepsilon^2/4$.

If $|r - p(1-p)| \leq \varepsilon^2/4$, then, by considering the off-diagonal term, we have:

$$\| M(\mathrm{y}) - M_0 \|_\infty \geq |\alpha - pp' + r|.$$

For $\delta' = r - p(1-p) \in [-\varepsilon^2/4, \varepsilon^2/4]$, using (4.126), we get that:

$$\begin{aligned}\alpha - pp' + r &= \alpha - p(1-p-\delta) + p(1-p) + \delta' \\ &\geq (1-\varepsilon)\varepsilon - \delta(1-\varepsilon-p) + \delta' \\ &\geq (1-\varepsilon)\varepsilon - (1-2\varepsilon)\varepsilon/2 - \varepsilon^2/4 \geq \varepsilon^2/4.\end{aligned}$$

We deduce that (4.123) holds if $|1-p-p'| \leq \varepsilon/2$ and $M_0 \in \mathcal{F}_-$.

Conclusion

Since (4.123) holds when $M_0 \in \mathcal{F}_+$, when $M_0 \in \mathcal{F}_-$ and either $|1-p-p'| > \varepsilon/2$ or $|1-p-p'| \leq \varepsilon/2$, we deduce that (4.123) holds under the condition of Lemma 4.41. □

Lemma 4.42. *Let $\varepsilon \in (0, 1/2)$, $\delta \in [-\varepsilon/2, \varepsilon/2]$, $f, g \in \mathcal{B}([0,1])$ such that $0 \leq f, g \leq 1-\varepsilon$ and $\int_{[0,1]}(f+g) = 1-\delta$. Then we have $\int_{[0,1]} fg \geq (1-\varepsilon)(\varepsilon-\delta)$, and this lower bound is sharp.*

Proof. Set $f_1 = \min(f,g)$ and $g_1 = \max(f,g)$ so that $0 \leq f_1 \leq g_1 \leq 1-\varepsilon$ and $\int_{[0,1]}(f_1+g_1) = 1-\delta$ and $\int_{[0,1]} f_1 g_1 = \int_{[0,1]} fg$. Set $h = \min(f_1, (1-\varepsilon-g_1))$ as well as $f_2 = f_1 - h$ and $g_2 = g_1 + h$ so that $0 \leq f_2 \leq g_2 \leq 1-\varepsilon$, $\int_{[0,1]}(f_2+g_2) = 1-\delta$ and

$$\int_{[0,1]} f_2\, g_2 = \int_{[0,1]} (f_1 - h)\,(g_1 + h) = \int_{[0,1]} f_1\, g_1 - \int_{[0,1]} (h(g_1 - f_1) + h^2) \leq \int_{[0,1]} f_1\, g_1 = \int_{[0,1]} fg.$$

Since by construction either $f_2(x) = 0$ or $g_2(x) = 1-\varepsilon$, we deduce that:

$$\int_{[0,1]} fg \geq \int_{[0,1]} f_2\, g_2 \geq (1-\varepsilon) \int_{[0,1]} f_2 = (1-\varepsilon)\left(1 - \delta - \int_{[0,1]} g_2\right) \geq (1-\varepsilon)(\varepsilon - \delta).$$

To see this lower bound is sharp, consider $g = 1-\varepsilon$ and $f = \varepsilon - \delta$. □

4.11.2 Proof of Proposition 4.28

We set:

$$\hat{Z}_n = (n-1)^{-1/2} M(\mathrm{x})^{-1/2} (\mathcal{D}^{(n+1)} - \mu(\mathrm{x})),$$

which is, conditionally on $\{X_{[2]} = \mathrm{x}\}$, distributed as the normalized and centered sum of $n-1$ independent random variables distributed as $Y = (Y_1, Y_2)$, with Y_1 and Y_2 Bernoulli random variables such that $\mathbb{E}[Y] = \mu(\mathrm{x})/(n-1)$ and $\mathrm{Cov}(Y,Y) = M(\mathrm{x})$.

Using Theorem 3.5 from [48] or Theorem 1.1 from [20], we get that:

$$\sup_{K \in \mathcal{K}} \left| \mathbb{P}\left(\hat{Z}_n \in K \middle| X_{[2]} = \mathrm{x} \right) - \mathbb{P}\left(Z \in K \right) \right| \leq 115\sqrt{2}\,\gamma,$$

where

$$\gamma = (n-1)\mathbb{E}\left[\left| (n-1)^{-1/2} M(\mathrm{x})^{-1/2} (Y - \mathbb{E}[Y]) \right|^3 \right].$$

Let $\|\cdot\|_\infty$ denote the matrix norm on the set $\mathbb{M}_2$ of real matrices of dimension 2×2 induced by the maximum vector norm on $\mathbb{R}^2$, which is the maximum absolute row sum:

$$\|M\|_\infty = \max_{1 \leq i \leq 2} \sum_{j=1}^{2} |M(i,j)|, \quad \text{for all } M \in \mathbb{M}_2. \tag{4.127}$$

Recall that $\|\cdot\|_\infty$ is an induced norm (that is $\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty$). For $M \in \mathbb{M}_2$ and $\mathrm{x} \in \mathbb{R}^2$, we have $|M\mathrm{x}| \leq \sqrt{2}\,\|M\|_\infty\,|\mathrm{x}|$. If $M \in \mathbb{M}_2$ is symmetric positive definite (which is only used for the second inequality and the equality), we get:

$$\|M\|_\infty^{1/2} \leq \|M^{1/2}\|_\infty \leq \sqrt{2}\,\|M\|_\infty^{1/2} \quad \text{and} \quad \|M^{-1}\|_\infty = \frac{\|M\|_\infty}{|\det(M)|}. \tag{4.128}$$

We deduce that if $M \in \mathbb{M}_2$ is symmetric positive definite, then

$$\|M^{-1/2}\|_\infty \leq \sqrt{2}|\det(M)|^{-1/2}\,\|M\|_\infty^{1/2}.$$

We obtain that for $n \geq 2$:

$$\begin{aligned}
\gamma &\leq 2^3 (n-1)^{-1/2}\,\|M(\mathrm{x})\|_\infty^{3/2} \det(M(\mathrm{x}))^{-3/2}\,\mathbb{E}\left[|Y - \mathbb{E}[Y]|^3 \right] \\
&\leq 2^{5/2} n^{-1/2} \det(M(\mathrm{x}))^{-3/2}\,\mathbb{E}\left[|Y_1 - \mathbb{E}[Y_1]|^3 + |Y_2 - \mathbb{E}[Y_2]|^3 \right] \\
&\leq 2^{3/2} n^{-1/2} \det(M(\mathrm{x}))^{-3/2},
\end{aligned}$$

where we used that $\|M(\mathrm{x})\|_\infty \leq 1/2$ for the second inequality and the convex inequality $(x+y)^p \leq 2^{p-1}(x^p + y^p)$ for the third and that $|Y_i - \mathbb{E}[Y_i]| \leq 1$ so that $\mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^3] \leq \mathrm{Var}\,(Y_i) \leq 1/4$. We deduce from Lemma 4.41 that there exists $C_0 > 0$ such that for all $\mathrm{x} = (x_1, x_2) \in [0,1]^2$ with $x_1 \neq x_2$ and all $n \geq 2$:

$$\sup_{K \in \mathcal{K}} \left| \mathbb{P}\left(\hat{Z}_n \in K \middle| X_{[2]} = \mathrm{x} \right) - \mathbb{P}\left(Z \in K \right) \right| \leq C_0\, n^{-1/2}.$$

To conclude, replace the convex set K in this formula by the convex set $\frac{M(\mathrm{x})^{-\frac{1}{2}}}{\sqrt{n-1}}(K - \mu(\mathrm{x}))$.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] R. Abraham and J.-F. Delmas. The forest associated with the record process on a Lévy tree. *Stochastic Process. Appl.*, 123(9):3497 – 3517, 2013.

[2] R. Abraham and J.-F. Delmas. Record process on the continuum random tree. *ALEA Lat. Am. J. Probab. Math. Stat.*, 10(1):225–251, 2013.

[3] R. Abraham and J.-F. Delmas. Local limits of conditioned Galton-Watson trees: the condensation case. *Electron. J. Probab.*, 19:no. 56, 29, 2014.

[4] R. Abraham and J.-F. Delmas. Local limits of conditioned Galton-Watson trees: the infinite spine case. *Electron. J. Probab.*, 19:no. 2, 19, 2014.

[5] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[6] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Modern Phys.*, 74(1):47–97, 2002.

[7] D. Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1(2):228–266, 1991.

[8] D. Aldous. The continuum random tree. I. *Ann. Probab.*, 19(1):1–28, 1991.

[9] D. Aldous. The continuum random tree. II. An overview. In *Stochastic analysis (Durham, 1990)*, volume 167 of *London Math. Soc. Lecture Note Ser.*, pages 23–70. Cambridge Univ. Press, Cambridge, 1991.

[10] D. Aldous. The continuum random tree. III. *Ann. Probab.*, 21(1):248–289, 1993.

[11] D. Aldous. Probability distributions on cladograms. In *Random discrete structures (Minneapolis, MN, 1993)*, volume 76 of *IMA Vol. Math. Appl.*, pages 1–18. Springer, New York, 1996.

[12] D. J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.

[13] D. J. Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.

[14] C. J. Anderson, S. Wasserman, and B. Crouch. A p* primer: Logit models for social networks. *Social Networks*, 21(1):37–66, 1999.

[15] K. B. Athreya and P. E. Ney. *Branching processes*. Springer-Verlag, New York-Heidelberg, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 196.

[16] N. Bacaër. *A short history of mathematical population dynamics*. Springer-Verlag London, Ltd., London, 2011.

[17] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[18] A. D. Barbour, M. Karoński, and A. Ruciński. A central limit theorem for decomposable random variables with applications to random graphs. *J. Combin. Theory Ser. B*, 47(2):125–145, 1989.

[19] I. Benjamini and O. Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:no. 23, 13, 2001.

[20] V. Bentkus. On the dependence of the Berry-Essèen bound on dimension. *J. Statist. Plann. Inference*, 113(2):385–402, 2003.

[21] J. Bertoin. *Lévy Processes*. Cambridge University Press, 1996.

[22] R. N. Bhattacharya and R. R. Rao. *Normal approximation and asymptotic expansions*, volume 64 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2010.

[23] P. Biane. Relations entre pont et excursion du mouvement brownien réel. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(1):1–7, 1986.

[24] P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 2009.

[25] P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.

[26] J. Blitzstein and P. Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.*, 6(4):489–522, 2010.

[27] M. Blum, O. François, and S. Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Ann. Appl. Probab.*, 16(4):2195–2214, 2006.

[28] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J. Combin.*, 1(4):311–316, 1980.

[29] B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

[30] B. Bollobás. The Erdős-Rényi theory of random graphs. In *Paul Erdős and his mathematics, II (Budapest, 1999)*, volume 11 of *Bolyai Soc. Math. Stud.*, pages 79–134. János Bolyai Math. Soc., Budapest, 2002.

[31] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures Algorithms*, 31(1):3–122, 2007.

[32] B. Bollobás and O. Riordan. Metrics for sparse graphs. In *Surveys in combinatorics 2009*, volume 365 of *London Math. Soc. Lecture Note Ser.*, pages 211–287. Cambridge Univ. Press, Cambridge, 2009.

[33] B. Bollobás and O. Riordan. Sparse graphs: metrics and random models. *Random Structures Algorithms*, 39(1):1–38, 2011.

[34] C. Borgs, J. Chayes, and D. Gamarnik. Convergent sequences of sparse graphs: a large deviations approach. *Random Structures Algorithms*, 51(1):52–89, 2017.

[35] C. Borgs, J. Chayes, J. Kahn, and L. Lovász. Left and right convergence of graphs with bounded degree. *Random Structures Algorithms*, 42(1):1–28, 2013.

[36] C. Borgs, J. Chayes, and L. Lovász. Moments of two-variable functions and the uniqueness of graph limits. *Geom. Funct. Anal.*, 19(6):1597–1619, 2010.

[37] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi. Limits of randomly grown graph sequences. *European J. Combin.*, 32(7):985–999, 2011.

[38] C. Borgs, J. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Counting graph homomorphisms. In *Topics in discrete mathematics*, volume 26 of *Algorithms Combin.*, pages 315–371. Springer, Berlin, 2006.

[39] C. Borgs, J. Chayes, and A. Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015.

[40] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, 219(6):1801–1851, 2008.

[41] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. of Math. (2)*, 176(1):151–219, 2012.

[42] G. Cardona, A. Mir, and F. Rosselló. Exact formulas for the variance of several balance indices under the Yule model. *J. Math. Biol.*, 67(6-7):1833–1846, 2013.

[43] A. Cayley. *The collected mathematical papers. Volume 8.* Cambridge Library Collection. Cambridge University Press, Cambridge, 2009. With a prefatory note by A. R. Forsyth, Reprint of the 1895 original.

[44] S. Chatterjee. An introduction to large deviations for random graphs. *Bull. Amer. Math. Soc. (N.S.)*, 53(4):617–642, 2016.

[45] S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *Ann. Appl. Probab.*, 21(4):1400–1435, 2011.

[46] S. Chatterjee and S. R. S. Varadhan. The large deviation principle for the Erdős-Rényi random graph. *European J. Combin.*, 32(7):1000–1017, 2011.

[47] B. Chauvin, J. Clément, and D. Gardy. *Arbres pour l'algorithmique.* 2018.

[48] L. H. Chen and X. Fang. Multivariate normal approximation by Stein's method: The concentration inequality approach. *arXiv preprint arXiv:1111.4073v2*, 2015.

[49] F. Chung, F. R. Chung, F. C. Graham, L. Lu, K. F. Chung, et al. *Complex graphs and networks.* Number 107. American Mathematical Soc., 2006.

[50] D. H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool*, 31:100–104, 1982.

[51] M. Coulson, R. E. Gaunt, and G. Reinert. Poisson approximation of subgraph counts in stochastic block models and a graphon model. *ESAIM Probab. Stat.*, 20:131–142, 2016.

[52] P. J. Davis. *Interpolation and approximation*. Dover Publications, Inc., New York, 1975.

[53] J.-F. Delmas, J.-S. Dhersin, and M. Sciauveau. Asymptotic for the cumulative distribution function of the degrees and homomorphism densities for random graphs sampled from a graphon. *arXiv preprint arXiv:1807.09989*, 2018.

[54] J.-F. Delmas, J.-S. Dhersin, and M. Sciauveau. Cost functionals for large (uniform and simply generated) random trees. *Electron. J. Probab.*, 23:1–36, 2018.

[55] L. Devroye. A note on the height of binary search trees. *J. Assoc. Comput. Mach.*, 33(3):489–498, 1986.

[56] L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Inform.*, 24(3):277–298, 1987.

[57] L. Devroye. Limit laws for local counters in random binary search trees. *Random Structures Algorithms*, 2(3):303–315, 1991.

[58] L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 249–314. Springer, Berlin, 1998.

[59] L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. In *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms Combin.*, pages 249–314. Springer, Berlin, 1998.

[60] L. Devroye. Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.*, 32(1):152–171, 2003.

[61] P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1):33–61, 2008.

[62] R. P. Dobrow and J. A. Fill. On the Markov chain for the move-to-root rule for binary search trees. *Ann. Appl. Probab.*, 5(1):1–19, 1995.

[63] R. P. Dobrow and J. A. Fill. Rates of convergence for the move-to-root Markov chain for binary search trees. *Ann. Appl. Probab.*, 5(1):20–36, 1995.

[64] R. P. Dobrow and J. A. Fill. Multiway trees of maximum and minimum probability under the random permutation model. *Combin. Probab. Comput.*, 5(4):351–371, 1996.

[65] R. P. Dobrow and J. A. Fill. Total path length for random recursive trees. *Combin. Probab. Comput.*, 8(4):317–333, 1999. Random graphs and combinatorial structures (Oberwolfach, 1997).

[66] A. A. Dobrynin, R. Entringer, and I. Gutman. Wiener index of trees: theory and applications. *Acta Appl. Math.*, 66(3):211–249, 2001.

[67] A. Dress, V. Moulton, and W. Terhalle. T-theory: an overview. *European J. Combin.*, 17(2-3):161–175, 1996. Discrete metric spaces (Bielefeld, 1994).

[68] M. Drmota. *Random trees*. SpringerWienNewYork, Vienna, 2009. An interplay between combinatorics and probability.

[69] T. Duquesne. A limit theorem for the contour process of conditioned Galton-Watson trees. *Ann. Probab.*, 31(2):996–1027, 2003.

[70] T. Duquesne and J.-F. Le Gall. Random trees, Lévy processes and spatial branching processes. *Astérisque*, (281):vi+147, 2002.

[71] T. Duquesne and J.-F. Le Gall. Probabilistic and fractal aspects of Lévy trees. *Probab. Theory Related Fields*, 131(4):553–603, 2005.

[72] R. Durrett. *Random graph dynamics*, volume 20 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2010.

[73] G. Elek and B. Szegedy. A measure-theoretic approach to the theory of dense hypergraphs. *Adv. Math.*, 231(3-4):1731–1772, 2012.

[74] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[75] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[76] P. Erdős and A. Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist.*, 38:343–347, 1961.

[77] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar.*, 12:261–267, 1961.

[78] S. N. Evans. *Probability and real trees*, volume 1920 of *Lecture Notes in Mathematics*. Springer, Berlin, 2008. Lectures from the 35th Summer School on Probability Theory held in Saint-Flour, July 6–23, 2005.

[79] S. N. Evans, J. Pitman, and A. Winter. Rayleigh processes, real trees, and root growth with re-grafting. *Probab. Theory Related Fields*, 134(1):81–126, 2006.

[80] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[81] X. Fang and A. Röllin. Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics. *Bernoulli*, 21(4):2157–2189, 2015.

[82] J. Felsenstein and J. Felenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

[83] V. Féray, P.-L. Méliot, and A. Nikeghbali. *Mod-ϕ convergence*. SpringerBriefs in Probability and Mathematical Statistics. Springer, 2016.

[84] V. Féray, P.-L. Méliot, and A. Nikeghbali. Graphons, permutons and the Thoma simplex: three mod-Gaussian moduli spaces. *arXiv preprint arXiv:1712.06841*, 2017.

[85] J. A. Fill. On the distribution of binary search trees under the random permutation model. *Random Struct. Algo.*, 8(1):1–25, 1996.

[86] J. A. Fill, P. Flajolet, and N. Kapur. Singularity analysis, Hadamard products, and tree recurrences. *J. Comput. Appl. Math.*, 174(2):271–313, 2005.

[87] J. A. Fill and S. Janson. Smoothness and decay properties of the limiting Quicksort density function. In *Mathematics and computer science (Versailles, 2000)*, Trends Math., pages 53–64. Birkhäuser, Basel, 2000.

[88] J. A. Fill and S. Janson. Precise logarithmic asymptotics for the right tails of some limit random variables for random trees. *Ann. Comb.*, 12(4):403–416, 2009.

[89] J. A. Fill and N. Kapur. Limiting distributions for additive functionals on Catalan trees. *Theoret. Comput. Sci.*, 326(1-3):69–102, 2004.

[90] J. A. Fill and N. Kapur. A repertoire for additive functionals of uniformly distributed m-ary search trees (extended abstract). In *2005 International Conference on Analysis of Algorithms*, Discrete Math. Theor. Comput. Sci. Proc., AD, pages 105–114 (electronic). Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2005.

[91] J. A. Fill and N. Kapur. Transfer theorems and asymptotic distributional results for m-ary search trees. *Random Struct. Algo.*, 26(4):359–391, 2005.

[92] P. Flajolet, X. Gourdon, and C. Martínez. Patterns in random binary search trees. *Random Struct. Algo.*, 11(3):223–244, 1997.

[93] D. J. Ford. *Probabilities on cladograms: Introduction to the alpha model.* ProQuest LLC, Ann Arbor, MI, 2006. Thesis (Ph.D.)–Stanford University.

[94] O. Frank and D. Strauss. Markov graphs. *J. Amer. Statist. Assoc.*, 81(395):832–842, 1986.

[95] M. Freedman, L. Lovász, and A. Schrijver. Reflection positivity, rank connectivity, and homomorphism of graphs. *J. Amer. Math. Soc.*, 20(1):37–51, 2007.

[96] A. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[97] C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43(6):2624–2652, 2015.

[98] W. Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. and Phys.*, 38:77–81, 1959/60.

[99] E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 1959.

[100] J. Gilmer and S. Kopparty. A local central limit theorem for triangles in a random graph. *Random Structures Algorithms*, 48(4):732–750, 2016.

[101] I. Gutman, S. Klavžar, and B. Mohar. *Fifty years of the Wiener index.* University, Department of Mathematics, 1997.

[102] T. E. Harris. *The theory of branching processes.* Dover Phoenix Editions. Dover Publications, Inc., Mineola, NY, 2002. Corrected reprint of the 1963 original [Springer, Berlin; MR0163361 (29 #664)].

[103] H. Hatami, L. Lovász, and B. Szegedy. Limits of locally-globally convergent graph sequences. *Geom. Funct. Anal.*, 24(1):269–296, 2014.

[104] S. B. Heard. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46(6):1818–1826, 1992.

[105] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948.

[106] P. W. Holland, K. B. Laskey, S. Leinhardt, and and. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.

[107] P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.*, 76(373):33–65, 1981. With comments by Ronald L. Breiger, Stephen E. Fienberg, Stanley Wasserman, Ove Frank and Shelby J. Haberman and a reply by the authors.

[108] C. Holmgren and S. Janson. Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.*, 20:no. 4, 51, 2015.

[109] C. Holmgren and S. Janson. Fringe trees, Crump-Mode-Jagers branching processes and m-ary search trees. *Probab. Surv.*, 14:53–154, 2017.

[110] D. N. Hoover. Row-column exchangeability and a generalized model for probability. In *Exchangeability in probability and statistics (Rome, 1981)*, pages 281–291. North-Holland, Amsterdam-New York, 1982.

[111] H.-K. Hwang and R. Neininger. Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM J. Comput.*, 31(6):1687–1722 (electronic), 2002.

[112] K. Itô and H. P. McKean, Jr. *Diffusion processes and their sample paths.* Springer-Verlag, Berlin-New York, 1974. Second printing, corrected, Die Grundlehren der mathematischen Wissenschaften, Band 125.

[113] S. Janson. The Wiener index of simply generated random trees. *Random Struct. Algo.*, 22(4):337–358, 2003.

[114] S. Janson. Simply generated trees, conditioned Galton-Watson trees, random allocations and condensation. *Probab. Surv.*, 9:103–252, 2012.

[115] S. Janson. Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton-Watson trees. *Random Struct. Algo.*, 48(1):57–101, 2016.

[116] S. Janson and P. Chassaing. The center of mass of the ISE and the Wiener index of trees. *Electron. Comm. Probab.*, 9:178–187 (electronic), 2004.

[117] S. Janson, T. Ł uczak, and A. Rucinski. *Random graphs.* Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[118] S. Janson and K. Nowicki. The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probab. Theory Related Fields*, 90(3):341–375, 1991.

[119] T. Jonsson and S. O. Stefánsson. Condensation in nongeneric trees. *J. Stat. Phys.*, 142(2):277–313, 2011.

[120] D. G. Kendall. The genealogy of genealogy: branching processes before (and after) 1873. *Bull. London Math. Soc.*, 7(3):225–253, 1975. With a French appendix containing Bienaymé's paper of 1845.

[121] D. P. Kennedy. The Galton-Watson process conditioned on the total progeny. *J. Appl. Probability*, 12(4):800–806, 1975.

[122] H. Kesten. Subdiffusive behavior of random walk on a random cluster. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(4):425–487, 1986.

[123] M. Kirxpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1993.

[124] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.*, 45(1):316–354, 2017.

[125] O. Klopp and N. Verzelen. Optimal graphon estimation in cut distance. *arXiv preprint arXiv:1703.05101*, 2017.

[126] D. E. Knuth. *The art of computer programming. Vol. 1.* Addison-Wesley, Reading, MA, 1997. Fundamental algorithms, Third edition [of MR0286317].

[127] D. E. Knuth. *The art of computer programming. Vol. 2.* Addison-Wesley, Reading, MA, 1998. Seminumerical algorithms, Third edition [of MR0286318].

[128] D. E. Knuth. *The art of computer programming. Vol. 3.* Addison-Wesley, Reading, MA, 1998. Sorting and searching, Second edition [of MR0445948].

[129] I. Kortchemski. Invariance principles for Galton-Watson trees conditioned on the number of leaves. *Stochastic Process. Appl.*, 122(9):3126–3172, 2012.

[130] I. Kortchemski. A simple proof of Duquesne's theorem on contour processes of conditioned Galton-Watson trees. In *Séminaire de Probabilités XLV*, volume 2078 of *Lecture Notes in Math.*, pages 537–558. Springer, Cham, 2013.

[131] I. Kortchemski. Limit theorems for conditioned non-generic Galton-Watson trees. *Ann. Inst. Henri Poincaré Probab. Stat.*, 51(2):489–511, 2015.

[132] K. Krokowski and C. Thaele. Multivariate central limit theorems for rademacher functionals with applications. *arXiv preprint arXiv:1701.07365*, 2017.

[133] P. Latouche and S. Robin. Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Stat. Comput.*, 26(6):1173–1185, 2016.

[134] J.-F. Le Gall. The uniform random tree in a Brownian excursion. *Probab. Theory Related Fields*, 96(3):369–383, 1993.

[135] J.-F. Le Gall. Random trees and applications. *Probab. Surv.*, 2:245–311, 2005.

[136] J.-F. Le Gall and Y. Le Jan. Branching processes in Lévy processes: the exploration process. *Ann. Probab.*, 26(1):213–252, 1998.

[137] J.-F. Le Gall and G. Miermont. Scaling limits of random trees and planar maps. In *Probability and statistical physics in two and more dimensions*, volume 15 of *Clay Math. Proc.*, pages 155–211. Amer. Math. Soc., Providence, RI, 2012.

[138] L. Lovász. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012.

[139] L. Lovász and B. Szegedy. Limits of dense graph sequences. *J. Combin. Theory Ser. B*, 96(6):933–957, 2006.

[140] L. M. Lovász. A short proof of the equivalence of left and right convergence for sparse graphs. *European J. Combin.*, 53:1–7, 2016.

[141] H. M. Mahmoud. Limiting distributions for path lengths in recursive trees. *Probab. Engrg. Inform. Sci.*, 5(1):53–59, 1991.

[142] H. M. Mahmoud. *Evolution of random search trees*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York, 1992. A Wiley-Interscience Publication.

[143] P.-A. G. Maugis, C. E. Priebe, S. C. Olhede, and P. J. Wolfe. Statistical inference for network samples using subgraph counts. *arXiv preprint arXiv:1701.00505*, 2017.

[144] A. Meir and J. W. Moon. On the altitude of nodes in random trees. *Canad. J. Math.*, 30(5):997–1015, 1978.

[145] A. Meir and J. W. Moon. On the log-product of the subtree-sizes of random trees. *Random Structures Algorithms*, 12(2):197–212, 1998.

[146] A. Mir, F. Rosselló, and L. Rotger. A new balance index for phylogenetic trees. *Math. Biosci.*, 241(1):125–136, 2013.

[147] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[148] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, probability and computing*, 7(3):295–305, 1998.

[149] A. O. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *The quarterly review of Biology*, 72(1):31–54, 1997.

[150] S. V. Nagaev, V. I. Chebotarev, and A. Y. Zolotukhin. A non-uniform bound of the remainder term in the central limit theorem for Bernoulli random variables. *J. Math. Sci. (N.Y.)*, 214(1):83–100, 2016.

[151] R. Neininger. On binary search tree recursions with monomials as toll functions. *J. Comput. Appl. Math.*, 142(1):185–196, 2002.

[152] R. Neininger. The Wiener index of random trees. *Combin. Probab. Comput.*, 11(6):587–597, 2002.

[153] J. Neveu. Arbres et processus de Galton-Watson. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(2):199–207, 1986.

[154] M. Newman. *Networks: an introduction*. Oxford university press, 2010.

[155] M. Newman, A.-L. Barabasi, and D. J. Watts. *The structure and dynamics of networks*, volume 19. Princeton University Press, 2011.

[156] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

[157] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[158] M. E. J. Newman. The structure and function of networks. In *Proceedings of the Europhysics Conference on Computational Physics (CCP 2001) (Aachen)*, volume 147, pages 40–45, 2002.

[159] K. Nowicki. Asymptotic normality of graph statistics. *J. Statist. Plann. Inference*, 21(2):209–222, 1989.

[160] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.*, 96(455):1077–1087, 2001.

[161] K. Nowicki and J. C. Wierman. Subgraph counts in random graphs using incomplete U-statistics methods. In *Proceedings of the First Japan Conference on Graph Theory and Applications (Hakone, 1986)*, volume 72, pages 299–310, 1988.

[162] J. Park and M. E. J. Newman. Statistical mechanics of networks. *Phys. Rev. E (3)*, 70(6):066117, 13, 2004.

[163] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[164] D. J. D. S. Price. Networks of scientific papers. *Science*, pages 510–515, 1965.

[165] D. Ralaivaosaona and S. Wagner. Additive functionals of d-ary increasing trees. *arXiv preprint arXiv:1605.03918*, 2016.

[166] M. Régnier. A limiting distribution for quicksort. *RAIRO Inform. Théor. Appl.*, 23(3):335–343, 1989.

[167] G. Reinert and A. Röllin. Random subgraph counts and U-statistics: multivariate normal approximation via exchangeable pairs and embedding. *J. Appl. Probab.*, 47(2):378–393, 2010.

[168] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[169] C. Richard. On q-functional equations and excursion moments. *Discrete Math.*, 309(1):207–230, 2009.

[170] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[171] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Itô calculus, Reprint of the second (1994) edition.

[172] U. Rösler. A limit theorem for "Quicksort". *RAIRO Inform. Théor. Appl.*, 25(1):85–100, 1991.

[173] U. Rösler and L. Rüschendorf. The contraction method for recursive algorithms. *Algorithmica*, 29(1-2):3–33, 2001.

[174] A. Ruciński. When are small subgraphs of a random graph normally distributed? *Probab. Theory Related Fields*, 78(1):1–10, 1988.

[175] M. Sackin. "good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972.

[176] R. Sedgewick. *Algorithms*. Addison-Wesley Series in Computer Science. Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1983.

[177] R. Sedgewick and P. Flajolet. *An introduction to the analysis of algorithms*. Pearson Education India, 2013.

[178] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[179] K.-T. Shao and R. R. Sokal. Tree balance. *Systematic Zoology*, 39(3):266–276, 1990.

[180] D. Strauss. On a general class of models for interaction. *SIAM Rev.*, 28(4):513–527, 1986.

[181] B. Stufler. The continuum random tree is the scaling limit of unlabelled unrooted trees. *arXiv preprint arXiv:1412.6333*, 2014.

[182] E. Szemerédi. Regular partitions of graphs. In *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*, volume 260 of *Colloq. Internat. CNRS*, pages 399–401. CNRS, Paris, 1978.

[183] L. Takács. On the total heights of random rooted binary trees. *J. Combin. Theory Ser. B*, 61(2):155–166, 1994.

[184] G. Tinhofer. Generating graphs uniformly at random. In *Computational graph theory*, volume 7 of *Comput. Suppl.*, pages 235–255. Springer, Vienna, 1990.

[185] N. Trinajstic. *Chemical graph theory*. Routledge, 2018.

[186] J. V. Uspensky. *Introduction to Mathematical Probability*. McGraw-Hill Book Company, New York, 1937.

[187] R. van der Hofstad. *Random graphs and complex networks. Vol. 1*. Cambridge Series in Statistical and Probabilistic Mathematics, [43]. Cambridge University Press, Cambridge, 2017.

[188] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[189] W. Vervaat. A relation between Brownian bridge and Brownian excursion. *Ann. Probab.*, 7(1):143–149, 1979.

[190] S. Wagner. Additive tree functionals with small toll functions and subtrees of random trees. In *23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12)*, Discrete Math. Theor. Comput. Sci. Proc., AQ, pages 67–80. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2012.

[191] S. Wagner. Central limit theorems for additive tree parameters with small toll functions. *Combin. Probab. Comput.*, 24(1):329–353, 2015.

[192] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[193] H. Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69(1):17–20, 1947.

[194] P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

[195] G. U. Yule et al. Ii.—a mathematical theory of evolution, based on the conclusions of dr. jc willis, fr s. *Phil. Trans. R. Soc. Lond. B*, 213(402-410):21–87, 1925.

[196] E. Zohoorian Azad. Asymptotic cost of cutting down random free trees. *Journal of The Iranian Statistical Society*, 11(1):57–73, 2012.